Model Comparison for Semantic Grouping

Francisco Vargas [1]   **Kamen Brestnichki** [1]   **Nils Hammerla** [1]

Abstract

We introduce a probabilistic framework for quantifying the semantic similarity between two groups of embeddings. We formulate the task of semantic similarity as a model comparison task in which we contrast a generative model which jointly models two sentences versus one that does not. We illustrate how this framework can be used for the Semantic Textual Similarity tasks using clear assumptions about how the embeddings of words are generated. We apply model comparison that utilises information criteria to address some of the shortcomings of Bayesian model comparison, whilst still penalising model complexity. We achieve competitive results by applying the proposed framework with an appropriate choice of likelihood on the STS datasets.

1. Introduction

The problem of Semantic Textual Similarity (STS), measuring how closely the meaning of one piece of text corresponds to that of another, has been studied in the hope of improving performance across various problems in Natural Language Processing (NLP), including information retrieval (Zheng & Callan, 2015). Recent progress in the learning of word embeddings (Mikolov et al., 2013b) has allowed the encoding of words using distributed vector representations, which capture semantic information through their location in the learned embedding space. Despite the extent to which semantic relations between words are captured in this space, it remains a challenge for researchers to adapt these individual word embeddings to express semantic similarity between word groups, like documents, sentences, and other textual formats.

Recent methods for STS rely on additive composition of word vectors (Arora et al., 2016; Blacoe & Lapata, 2012; Mitchell & Lapata, 2008; 2010; Wieting et al., 2015) or

---

[1]Babylon Health. Correspondence to: Francisco Vargas <francisco.vargas@babylonhealth.com>.

deep learning architectures (Kiros et al., 2015), both of which summarise a sentence through a single embedding. The resulting sentence vectors are then compared using cosine similarity — a choice stemming only from the fact that cosine similarity gives good empirical results. It is difficult for a practitioner to utilize word vectors efficiently, as the underlying assumptions in the similarity measure are not well understood and the assumptions made at the word vector level are not clearly defined. For example, embedding magnitude in (Arora et al., 2016) seems to matter at the word level, but not at the sentence level. That is, normalizing word embeddings before the SIF estimate decreases results considerably, while normalising sentence embeddings has no effect on results.

The main contribution of our work is the proposal of a framework that addresses these issues by explicitly deriving the similarity measure through a chosen generative model of embeddings. Via this design process, a practitioner can encode suitable assumptions and constraints that may be favourable to the application of interest. Furthermore, this framework puts forward a new research direction that could help improve the understanding of semantic similarity by allowing practitioners to study suitable embedding distributions and assess how these perform.

The second contribution of our work is the derivation of a similarity measure that performs well in an online setting. Online settings are both practical and key to use-cases that involve information retrieval in dialogue systems. For example, in a chat-bot application new queries will arrive one at a time and methods such as the one proposed in (Arora et al., 2016) will not perform as strongly as they do on the benchmark datasets. This is because one cannot perform the required data pre-processing on the entire query dataset, which will not be available a priori in online settings. Whilst our framework produces an online similarity metric, it remains competitive to offline methods such as (Arora et al., 2016). We achieve results comparable to (Arora et al., 2016) on the STS dataset in $\mathcal{O}(nd)$ time — compared to the $\mathcal{O}(nd^2)$ average complexity of their method (where n is the number of words in a sentence, and d is the embedding size).

2. Background

The compositional nature of distributed representations demonstrated in (Mikolov et al., 2013b) and (Pennington et al., 2014) indicate the presence of semantic groups in the representation space of word embeddings; an idea which has been further explored in (Athiwaratkun & Wilson, 2017). Under this assumption, the task of semantic similarity can be formulated as the following question: "Are the two sentences (as groups of words) samples from the same semantic group?". Utilising this rephrasing, this work formulates the task of semantic similarity between two arbitrary groups of objects as a model comparison problem. Taking inspiration from (Ghahramani & Heller, 2006) and (Marshall et al., 2006) we propose the generative models for groups (e.g. sentences) $\mathcal{D}_1, \mathcal{D}_2$ seen in Figure 1.

The Bayes Factor for this graphical model is then formally defined as

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = \log \frac{p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1)}{p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_2)}$$

$$= \log \frac{p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1)}{p(\mathcal{D}_1 | \mathcal{M}_2) p(\mathcal{D}_2 | \mathcal{M}_2)}. \quad (1)$$

To obtain the evidence $p(\mathcal{D} | \mathcal{M}_i)$ the parameters of the model must be marginalised

$$p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1) = \int p(\mathcal{D}_1, \mathcal{D}_2 | \theta) p(\theta) d\theta$$

$$p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1) = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_1 \oplus \mathcal{D}_2} p(\boldsymbol{w}_k | \theta) p(\theta) d\theta,$$

$$p(\mathcal{D}_i | \mathcal{M}_2) = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_i} p(\boldsymbol{w}_k | \theta) p(\theta) d\theta,$$

where $\oplus$ means concatenation and $\boldsymbol{w}_k$ is a word embedding.

Computing the semantic similarity score of the two groups $\mathcal{D}_1, \mathcal{D}_2$ under the Bayesian framework requires selecting
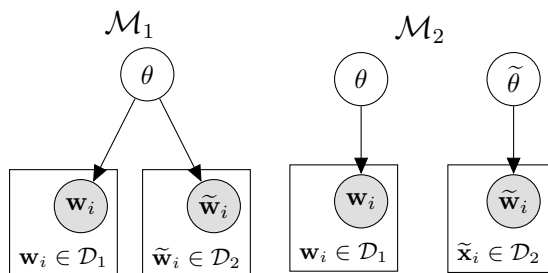

Figure 1. On the left, $\mathcal{M}_1$ assumes that both datasets are generated i.i.d. from the same parametric distribution. On the right, $\mathcal{M}_2$ assumes that the datasets are generated i.i.d. from distinct parametric distributions.

a reasonable model likelihood $p(\boldsymbol{w}_k | \theta)$, prior density on the parameters $p(\theta)$, and computing the marginal evidence specified above. Computing the evidence can be computationally intensive and usually requires approximation. What is more, the Bayes factor is very sensitive to the choice of prior and can result in estimates that heavily underfit the data (especially under a vague prior, as shown in Appendix E [1]), having the tendency to select the simpler model; which is argued further by (M. S. Bartlett, 1957) and (Akaike et al., 1981). This is handled in (Ghahramani & Heller, 2006) by using the empirical Bayes procedure, a shortcoming of which is the issue of double counting (Berger, 2000) and thus being prone to over-fitting. We address these issues by choosing to work with information criteria based model comparison as opposed to using the Bayes factor. The details are described in Section 3.

We are not aware of any prior work on sentence similarity that uses our approach. We employ a generative model for sentences similar to (Arora et al., 2016), but our contribution differs from theirs in that our similarity is based on the aforementioned model comparison test, whilst theirs is based on the inner product of sentence embeddings derived as maximum likelihood estimators. The method by (Marshall et al., 2006) applies the same score as in Equation 1 in the context of merging datasets whilst we focus on information retrieval and semantic textual similarity.

3. Methodology

We address the shortcomings of the Bayes Factor described in Section 2 by proposing model comparison criteria that minimise Kullback-Leibler (KL) divergence (denoted in equations as D_{KL}) across a candidate set of models. This results in a penalised likelihood ratio test which gives competitive results. This approach relies on Information Theoretic Criteria to design a log-ratio-like test that is prior free and robust to overfitting.

We seek to compare models $\mathcal{M}_1$ and $\mathcal{M}_2$ using Information Criteria (IC) to assess the goodness of fit of each model. There are multiple IC used for model selection, each with different settings to which they are better suited. The IC which we will be working with have the general form

$$\mathrm{IC}(\mathcal{D}, \mathcal{M}) = -\left(\alpha \mathcal{L}(\hat{\boldsymbol{\theta}} | \mathcal{D}, \mathcal{M}) + \Omega(\mathcal{D}, \mathcal{M}) \right),$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}} | \mathcal{D}, \mathcal{M}) = \sum_i^n \mathcal{L}(\hat{\boldsymbol{\theta}} | \boldsymbol{w}_i, \mathcal{M})$ is the maximised value of the log likelihood function for model $\mathcal{M}$, α is a scalar derived for each IC, and $\Omega(\mathcal{D}, \mathcal{M})$ represents a model complexity penalty term which is model and IC specific. Using this general formulation for the involved infor-

---

[1] All appendices are provided in the supplementary materials.

mation criteria yields the similarity score

$$
\begin{aligned}
\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) &= \\
&= -\operatorname{IC}(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_1) + \operatorname{IC}(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_2) \\
&= \alpha \Big(\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_{1,2} | \mathcal{D}, \mathcal{M}_1) - (\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_1 | \mathcal{D}_1, \mathcal{M}_2) + \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_2 | \mathcal{D}_1, \mathcal{M}_2)) \Big) \\
&\quad - \Omega\left(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_1\right) + \Omega\left(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_2\right).
\end{aligned}
$$

Tversky's contrast model[2] describes what a good similarity is from a cognitive science perspective. Interestingly, the semantic interpretation of the similarity we have derived is similar to the one described in that work. We want to contrast the commonalities of the two datasets (through shared parameters) to the distinctive features of each dataset (through independently fit parameters).

Examples of these criteria can be put into two broad classes. The Bayesian Information Criterion (BIC) is an example of an IC that approximates the model evidence directly, as defined in (Schwarz et al., 1978). Empirically it has been shown that the BIC is likely to underfit the data, especially when the number of samples is small (Dziak et al., 2012). We provide additional empirical results in Appendix E, showing BIC is not a good fit for the STS task as sentences contain a relatively small number of words (samples). Thus we focus on the second class - Information Theoretic Criteria.

3.1. Information Theoretic Criteria

The Information Theoretic Criteria (ITC) are a family of model selection criteria. The task they address is evaluating the expected quality of an estimated model specified by $\mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{w})$ when it is used to generate unseen data from the true distribution $G(\boldsymbol{w})$, as defined in (Konishi & Kitagawa, 2008b). This family of criteria perform this evaluation by using the KL divergence between the true model $G(\boldsymbol{w})$ and the fitted model $\mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{w})$, with the aim of selecting the model (from a given set of models) that minimizes the quantity

$$
\begin{aligned}
D_{KL}\left(G(\boldsymbol{w}) \middle\| p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right) &= \mathbb{E}_G\left[\ln \frac{G(\boldsymbol{w})}{p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})}\right] \\
&= -H_G(\boldsymbol{w}) - \mathbb{E}_G\left[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right].
\end{aligned}
$$

The entropy of the true model $H_G(\boldsymbol{w})$ will remain constant across different likelihoods. Thus, the quantity of interest in the definition of the information criterion under consideration is given by the expected log likelihood under the true model $\mathbb{E}_G[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})]$. The goal is to find a good estimator for this quantity and one such estimator is given by the

---

[2]As described in the paragraph surrounding Equation (9) of (Tenenbaum & Griffiths, 2001)

normalized maximum log likelihood

$$
\mathbb{E}_{\hat{G}}\left[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right] = \frac{1}{n}\sum_{i=1}^{n} \ln p(\boldsymbol{w}_i|\hat{\boldsymbol{\theta}}),
$$

where $\hat{G}$ represents the empirical distribution. This estimator introduces a bias that varies with respect to the dimension of the model's parameter vector $\boldsymbol{\theta}$ and requires a correction in order to carry out a fair comparison of information criteria between models. A model specific correction is derived resulting in the following IC, called Takeuchi Information Criterion (TIC) in (Takeuchi, 1976)

$$
\hat{\boldsymbol{J}} = -\frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w}_i)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},
$$

$$
\hat{\boldsymbol{\mathcal{I}}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w}_i)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w}_i)^\top\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},
$$

$$
\operatorname{TIC}(\mathcal{D}, \mathcal{M}) = -2\left(\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{D}, \mathcal{M}) - \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}\right)\right). \quad (2)
$$

For the case where we assume our model has the same parametric form as the true model and as $n \to \infty$, the equality $\hat{\boldsymbol{\mathcal{I}}} = \hat{\boldsymbol{J}}$ holds resulting in a penalty of $\operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}\right) = \operatorname{tr}(\boldsymbol{I}_k) = k$, where k is the number of model parameters, as shown in (Konishi & Kitagawa, 2008a). This results in the Akaike Information Criterion (AIC) (Akaike, 1974)

$$
\operatorname{AIC}(\mathcal{D}, \mathcal{M}) = -2(\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{D}, \mathcal{M}) - k).
$$

The AIC simplification of TIC relies on several assumptions that hold true in the big data limit. However, as shown in Appendix F, for models with a high number of parameters, TIC may prove unstable and thus AIC will generally perform better. In this study we consider and contrast both.

We show in Appendix A that under the TIC we have the following similarity (where we omit the conditioning on the models for brevity)

$$
\begin{aligned}
\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = 2\Big(& \mathcal{L}(\hat{\boldsymbol{\theta}}_{1,2}|\mathcal{D}_{1,2}) - \mathcal{L}(\hat{\boldsymbol{\theta}}_1|\mathcal{D}_1) - \mathcal{L}(\hat{\boldsymbol{\theta}}_2|\mathcal{D}_2) \\
& - \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_{1,2}\hat{\boldsymbol{J}}_{1,2}^{-1}\right) + \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_1\hat{\boldsymbol{J}}_1^{-1}\right) + \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_2\hat{\boldsymbol{J}}_2^{-1}\right) \Big),
\end{aligned}
$$

and therefore under the AIC we have the following similarity

$$
\begin{aligned}
\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = 2\Big(& \mathcal{L}(\hat{\boldsymbol{\theta}}_{1,2}|\mathcal{D}_{1,2}) - \mathcal{L}(\hat{\boldsymbol{\theta}}_1|\mathcal{D}_1) \\
& - \mathcal{L}(\hat{\boldsymbol{\theta}}_2|\mathcal{D}_2) + k \Big),
\end{aligned}
$$

where $\mathcal{D}_{1,2} = \mathcal{D}_1 \oplus \mathcal{D}_2$.

4. Word Embedding Likelihoods

In this section we will illustrate how to derive a similarity score under our ITC framework by choosing a likelihood function that incorporates our prior assumptions about the generating process of the data. Adopting the viewpoint of a practitioner, we would like to compare the performance of two models — one that ignores word embedding magnitude, and one that makes use of it. Our modelling choices for each assumption are the von Mises-Fisher (vMF) and Gaussian likelihoods respectively. The comparison between the two likelihoods we provide in Section 5 provides empirical evidence as to which approach is better suited to modelling word embeddings. As we will see the TIC penalties of both likelihoods we consider can be calculated in $O(nd)$, thus not increasing the time complexity of the algorithm.

4.1. Von Mises-Fisher Likelihood

Cosine similarity is often used to measure the semantic similarity of words in various information retrieval tasks. Thus, we want to explore a distribution induced by the cosine similarity measure. We model our embeddings as vectors lying on the surface of the $d-1$ dimensional unit hypersphere $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and i.i.d. according to a vMF likelihood (Fisher et al., 1993)

$$p(\boldsymbol{w}|\boldsymbol{\mu}, \kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right)$$

$$= \frac{1}{Z(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right),$$

where $\boldsymbol{\mu}$ is the mean direction vector and κ is the concentration, with supports $||\boldsymbol{\mu}|| = ||\boldsymbol{w}|| = 1$, $\kappa \geq 0$. The term $I_\nu(\kappa)$ Corresponds to a modified Bessel function of the first kind with order ν.

In this work we reparameterise the random variable to polar hypersphericals $\boldsymbol{w}(\boldsymbol{\phi})$ ($\boldsymbol{\phi} = (\phi_1, ..., \phi_{d-1})^\top$) and $\boldsymbol{\mu}(\boldsymbol{\theta})$ ($\boldsymbol{\theta} = (\theta_1, ..., \theta_{d-1})^\top$) as adopted in (Mabdia, 1975). Further details can be found in Appendix B.

We prove (in Appendix C) that the mixed derivatives of the vMF log likelihood are a constant (with respect to $\boldsymbol{\theta}$) times $\partial \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi})/\partial \theta_k$. Thus, evaluated at the MLE, these entries are zero. Thus, we know $\hat{\boldsymbol{J}} = \text{diag}(\hat{J}_{11}, ..., \hat{J}_{dd})$ and we can express the TIC penalty described in Equation 2 as

$$\text{tr}(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}) = \sum_{i=1}^d \hat{J}_{ii}^{-1} \hat{\mathcal{I}}_{ii} = \hat{J}_{11}^{-1} \left(\frac{\partial}{\partial \kappa} \mathcal{L}(\boldsymbol{\theta}, \kappa|\mathcal{D})\right)^2$$

$$+ \sum_{i=2}^d \hat{J}_{ii}^{-1} \left(\frac{\partial}{\partial \theta_{i-1}} \mathcal{L}(\boldsymbol{\theta}, \kappa|\mathcal{D})\right)^2. \tag{3}$$

This quantity only requires $\mathcal{O}(nd)$ operations to compute and thus does not increase the asymptotic complexity of the algorithm.

The closed form of the similarity measure for two sentences $\mathcal{D}_1, \mathcal{D}_2$ of length m and l respectively under this model is then

$$\begin{aligned}
\textbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = &(m+l)\hat{\kappa}_{1,2}\bar{R}_{1,2} - m\hat{\kappa}_1\bar{R}_1 - l\hat{\kappa}_2\bar{R}_2 \\
&- (m+l)\log Z(\hat{\kappa}_{1,2}) + m\log Z(\hat{\kappa}_1) + l\log Z(\hat{\kappa}_2) \\
&- \text{tr}(\hat{\boldsymbol{\mathcal{I}}}_{1,2}\hat{\boldsymbol{J}}_{1,2}^{-1}) + \text{tr}(\hat{\boldsymbol{\mathcal{I}}}_1\hat{\boldsymbol{J}}_1^{-1}) + \text{tr}(\hat{\boldsymbol{\mathcal{I}}}_2\hat{\boldsymbol{J}}_2^{-1}),
\end{aligned}$$

where the Jacobian terms (from the reparametrisation) cancel out. The subscripts indicate the sentence, with $1, 2$ meaning the concatenation of the two sentences.

4.2. Gaussian Likelihood

(Schakel & Wilson, 2015) show that some frequency information is contained in the magnitude of word embeddings. This motivates a choice of a likelihood function that is not constrained to the unit hypersphere and possibly the simplest such choice is the Gaussian likelihood. Due to the small size of sentences[3] we choose a diagonal covariance Gaussian models. The Gaussian likelihood is then

$$p(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}},$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix.

Our framework further allows us to compare two models and pick the better one without having access to a similarity corpus, as long as the comparison is done on the same data. As an example, we compare the diagonal Gaussian with a more restricted version of itself — the spherical Gaussian. We compute the average AIC on a corpus of sentences and observe that the average AIC of the diagonal Gaussian is lower than that of the spherical one. This suggests that the diagonal Gaussian better describes the distribution of word vectors in a sentence and thus will produce a better similarity. We confirm this result in Appendix H, where we also provide the average AIC scores for each model.

As with the vMF likelihood, we prove in Appendix D that the Hessian of the log likelihood evaluated at the MLE is diagonal. The TIC correction is a sum of $O(d)$ terms, similar to the form in Equation 3. This can be nicely written in terms of biased sample kurtosis (denoted $\hat{\boldsymbol{\kappa}}$)

$$\text{tr}(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}) = \frac{1}{2}\left(d + \sum_{i=1}^d \frac{(\hat{\boldsymbol{\mu}}_4)_i}{\hat{\sigma}_i^4}\right) = \frac{d}{2} + \sum_{i=1}^d \frac{\hat{\kappa}_i}{2}$$

The closed form of the similarity measure for two sentences $\mathcal{D}_1, \mathcal{D}_2$ of length m and l respectively under this model is

---

[3]The covariance matrix of n samples with d dimensions such that $n < d$ results in a low rank matrix.

Table 1. Comparison of Spearman correlations on the STS datasets between the two similarity measures we introduce in the text. The average is weighted according to dataset size.

Embedding	Method	STS12	STS13	STS14	STS15	STS16	Average
FastText	vMF+TIC	0.5219	0.5147	0.5719	0.6456	0.6347	0.5762
	Diag+AIC	**0.6193**	**0.6334**	**0.6721**	**0.7328**	**0.7518**	**0.6764**
GloVe	vMF+TIC	0.5421	0.5598	0.5736	0.6474	0.6168	0.5859
	Diag+AIC	**0.6031**	**0.6131**	**0.6445**	**0.7171**	**0.7346**	**0.6564**
Word2Vec GN	vMF+TIC	0.5665	0.5735	0.6062	0.6681	0.6510	0.6115
	Diag+AIC	**0.5957**	**0.6358**	**0.6614**	**0.7213**	**0.7187**	**0.6618**

then

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) =$$

$$= \sum_{i=1}^{d} -(m+l)\ln(\hat{\boldsymbol{\sigma}}_{1,2})_i + m\ln(\hat{\boldsymbol{\sigma}}_1)_i + l\ln(\hat{\boldsymbol{\sigma}}_2)_i +$$

$$+ \frac{d}{2} + \frac{1}{2}\sum_{i=1}^{d} -(\hat{\boldsymbol{\kappa}}_{1,2})_i + (\hat{\boldsymbol{\kappa}}_1)_i + (\hat{\boldsymbol{\kappa}}_2)_i$$

where the subscripts indicate the sentence, with $1, 2$ meaning the concatenation of the two sentences.

5. Experiments

We assess our methods' performance on the Semantic Textual Similarity (STS) datasets [4] (Agirre et al., 2012; 2013; 2014; 2015; 2016). The objective of these tasks is to estimate the similarity between two given sentences, validated against human scores. In our experiments, we assess on the pre-trained GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2016) , and Word2Vec GN (Mikolov et al., 2013a) word embeddings. For the vMF distribution, we normalize the word embeddings to be of length 1. Some of the sentences are left with a single word after querying the word embeddings, making the MLE of the κ parameter of the vMF and $\boldsymbol{\Sigma}$ parameter of the Gaussian ill-defined, which in turn causes the similarity metric to take on undefined values. We overcome this issue by padding each sentence with an arbitrary embedding of a word or punctuation symbol from the embedding lexicon (i.e. '.' or 'the'). Our code builds on top of SentEval (Conneau & Kiela, 2018) and is available at https://github.com/Babylonpartners/MCSG.

We first compare our methods: vMF likelihood with TIC correction (vMF+TIC) and diagonal Gaussian likelihood with AIC correction (Diag+AIC) against each other. Then, the better method is compared against mean word vector (MWV), word mover's distance (WMD) (Kusner et al.,

[4]The STS13 dataset does not include the proprietary SMT dataset that was available with the original release of STS.

2015) [5], smooth inverse frequency (SIF), and SIF with principal component removal as defined in (Arora et al., 2016) [6]. We re-ran these models under our experimental setup, to ensure a fair comparison. The metric used is the average Spearman correlation score over each dataset, weighted by the number of sentences. The choice of Spearman correlation is given by its non-parametric nature (assumes no distribution over the scores), as well as measuring any monotonic relationship between the two compared quantities.

5.1. Embedding Magnitude

A practitioner may want to learn more about a given set of word embeddings, and the way these embeddings were trained may not allow the user to understand the importance of certain features — say embedding magnitude. This is where our framework can be used to help build intuition by comparing a likelihood that implicitly incorporates embedding magnitude to one that does not.

We present the comparison between the similarities derived from the vMF and Gaussian likelihoods in Table 4.2. We note that the Gaussian is a much better modelling choice, beating the vMF on every dataset by a margin of at least 0.05 (5%) on average, with each of the three word embeddings. This is strong evidence that the information encoded in an embedding's magnitude is useful for tasks such as semantic similarity. This further motivates the conjecture that frequency information is contained in word embedding magnitude, as explored in (Schakel & Wilson, 2015).

5.2. Online Scenario

In an online setting, one cannot perform the principal component removal described in (Arora et al., 2016), as that pre-processing requires access to the entire query dataset a priori.

The comparison against the baseline methods are presented

[5]https://github.com/mkusner/wmd
[6]https://github.com/PrincetonML/SIF

Table 2. Comparison of Spearman correlations on the STS datasets between our best model (Diag+AIC) and SIF, WMD and MWV for three different word vectors. The average is weighted according to dataset size.

Embedding	Method	STS12	STS13	STS14	STS15	STS16	Average
FastText	Diag+AIC	**0.6193**	0.6334	**0.6721**	0.7328	**0.7518**	0.6764
	SIF	0.6079	**0.6989**	**0.6777**	**0.7436**	0.7135	**0.6821**
	MWV	0.5994	0.6494	0.6473	0.7114	0.6814	0.6542
	WMD	0.5576	0.5146	0.5915	0.6800	0.6402	0.5997
GloVe	Diag+AIC	**0.6031**	0.6131	**0.6445**	**0.7171**	**0.7346**	**0.6564**
	SIF	0.5774	**0.6319**	0.6135	0.6740	0.6589	0.6255
	MWV	0.5526	0.5643	0.5625	0.6314	0.5804	0.5784
	WMD	0.5516	0.5007	0.5811	0.6704	0.6246	0.5896
Word2Vec GN	Diag+AIC	**0.5957**	0.6358	**0.6614**	**0.7213**	**0.7187**	**0.6618**
	SIF	0.5697	**0.6594**	**0.6669**	**0.7261**	0.6952	**0.6588**
	MWV	0.5744	0.6330	0.6561	0.7040	0.6617	0.6451
	WMD	0.5554	0.5250	0.6074	0.6730	0.6399	0.6034

Table 3. Comparison of Spearman correlations on the STS datasets between our best model (Diag+AIC) and SIF+PCA for three different word vectors.

Word vectors	Method	STS12	STS13	STS14	STS15	STS16	Average
FastText	Diag+AIC	**0.6193**	0.6334	0.6721	0.7328	**0.7518**	0.6764
	SIF+PCA	0.5945	**0.7149**	**0.6824**	**0.7474**	0.7271	**0.6843**
GloVe	Diag+AIC	**0.6031**	0.6131	0.6445	**0.7171**	**0.7346**	0.6564
	SIF+PCA	0.5732	**0.6843**	**0.6546**	0.7166	0.6931	**0.6589**
Word2Vec GN	Diag+AIC	**0.5957**	0.6358	0.6614	0.7213	**0.7187**	0.6618
	SIF+PCA	0.5602	**0.6773**	**0.6722**	**0.7354**	0.7111	**0.6639**

in Table 5.2. The method proposed in this work is able to out-perform the standard weighting induced by MWV, as well as the WMD approach on all datasets, with each of the three word embeddings considered. We outperform SIF using the GloVe embeddings by 0.309 (3.09%) and effectively tie when using the FastText and Word2Vec GN embeddings with a difference of 0.0057 (0.57%) and 0.0030 (0.30%) respectively. In Appendix I we conduct a significance analysis. We show that SIF and our method are on par with each other and that both significantly outperform MWV and WMD when using the GloVe embeddings.

5.3. Offline Scenario

There are use-cases in which the entire dataset of sentences is available at evaluation time — for example, in clustering applications. For this scenario, we compare against the SIF weightings, augmented with the additional pre-processing technique seen in (Arora et al., 2016). We need only consider this baseline, as it outperforms all others by a large margin.

The results are shown in Table 5.2. We remain competitive with SIF+PCA on all three word embeddings, being able to match very closely on GloVe embeddings. On the FastText and Word2Vec embeddings, our method is less than 0.01 (1%) lower on average than SIF+PCA.

6. Conclusion

We've presented a new approach to similarity measurement that achieves competitive performance to standard methods in both online and offline settings. Our method requires a set of clear choices — model, likelihood and information criterion. From that, a comparison framework is naturally derived, which supplies us with a statistically justified similarity measure (by utilizing ITC to reduce the resulting model-comparison bias). This framework is suitable for a variety of modelling scenarios, due to the freedom in specifying the generative process. The graphical model we employ is adaptable to encode structural dependencies beyond the i.i.d. data-generating process we have assumed throughout this study — for example, an auto-regressive (sequential) model may be assumed if the practitioner suspects

that word order matters (i.e. compare "Does she want to get pregnant?" to "She does want to get pregnant.")

In this study, we conjecture that the von Mises-Fisher distribution lends itself to representing word embeddings well, if their magnitude is disregarded and a unimodal distribution over individual sentences is assumed. Relaxing the former assumption, we also model word embeddings with a Gaussian likelihood. As this improves results, it suggests that word embedding magnitude carries information relevant for sentence level tasks, which agrees with prior intuition built from (Schakel & Wilson, 2015). We hope that this framework could be a stepping stone in using more complex and accurate generative models of text to assess semantic similarity. For example, relaxing the assumption of unimodality is an interesting area for future research.

Acknowledgements

References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. 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*, pp. 385–393. Association for Computational Linguistics, 2012.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. * 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, pp. 32–43, 2013.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 81–91, 2014.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. 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)*, pp. 252–263, 2015.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511, 2016.

Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Akaike, H. et al. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):3–14, 1981.

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

Athiwaratkun, B. and Wilson, A. G. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017.

Berger, J. O. Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association*, 95(452):1269–1276, 2000. ISSN 01621459. URL http://www.jstor.org/stable/2669768.

Blacoe, W. and Lapata, M. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 546–556. Association for Computational Linguistics, 2012.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. URL http://arxiv.org/abs/1607.04606.

Conneau, A. and Kiela, D. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.

Dziak, J. J., Coffman, D. L., Lanza, S. T., and Li, R. Sensitivity and specificity of information criteria. *The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University*, 16(30):140, 2012.

Fisher, N. I., Lewis, T., and Embleton, B. J. *Statistical analysis of spherical data*. Cambridge university press, 1993.

Ghahramani, Z. and Heller, K. A. Bayesian sets. In *Advances in neural information processing systems*, pp. 435–442, 2006.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. Skip-thought vectors.

CoRR, abs/1506.06726, 2015. URL http://arxiv.org/abs/1506.06726.

Konishi, S. and Kitagawa, G. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008a.

Konishi, S. and Kitagawa, G. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008b.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/kusnerb15.html.

M. S. Bartlett, M. S. A comment on d. v. lindley's statistical paradox. *Biometrika*, 44(3/4):533–534, 1957.

Mabdia, K. Distribution theory for the von mises-fisher distribution and ite application. *Statistical Distributions for Scientific Work*, 1:113–30, 1975.

Marshall, P., Rajguru, N., and Slosar, A. Bayesian evidence as a tool for comparing datasets. *Physical Review D*, 73 (6):067302, 2006.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.

Mitchell, J. and Lapata, M. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pp. 236–244, 2008.

Mitchell, J. and Lapata, M. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

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

Schakel, A. M. and Wilson, B. J. Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297*, 2015.

Schwarz, G. et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Takeuchi, K. The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153:12–18, 1976.

Tenenbaum, J. B. and Griffiths, T. L. Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, 24(4):629–640, 2001.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

Zheng, G. and Callan, J. Learning to reweight terms with distributed representations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pp. 575–584, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. doi: 10.1145/2766462.2767700. URL http://doi.acm.org/10.1145/2766462.2767700.