Estimating Mutual Information Between Dense Word Embeddings

Vitalii Zhelezniak, Aleksandar Savkov & Nils Hammerla
Babylon Health
{firstname.lastname}@babylonhealth.com

Abstract

Word embedding-based similarity measures are currently among the top-performing methods on unsupervised semantic textual similarity (STS) tasks. Recent work has increasingly adopted a statistical view on these embeddings, with some of the top approaches being essentially various correlations (which include the famous cosine similarity). Another excellent candidate for a similarity measure is mutual information (MI), which can capture arbitrary dependencies between the variables and has a simple and intuitive expression. Unfortunately, its use in the context of dense word embeddings has so far been avoided due to difficulties with estimating MI for continuous data. In this work we go through a vast literature on estimating MI in such cases and single out the most promising methods, yielding a simple and elegant similarity measure for word embeddings. We show that mutual information is a viable alternative to correlations, gives an excellent signal that correlates well with human judgements of similarity and rivals existing state-of-the-art unsupervised methods.

1 Introduction

Neural text embeddings learned from unlabeled data are a key component of modern approaches to semantic textual similarity (STS). Despite the impressive performance of large pretrained models (Kiros et al., 2015; Conneau et al., 2017; Subramanian et al., 2018; Cer et al., 2018; Peters et al., 2018; Radford, 2018; Devlin et al., 2018; Dai et al., 2019; Yang et al., 2019a) on a a plethora of hard NLP tasks, deep models do not currently offer a clear advantage over much simpler static word embeddings (Bengio et al., 2003; Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Joulin et al., 2017) on standard unsupervised STS benchmarks (Hill et al., 2016; Arora et al., 2017; Wieting et al., 2016; Wieting and Gimpel, 2018; Zhelezniak et al., 2019b,a,c). Instead, the main sources of improvement here have come from training on supervised paraphrastic corpora (Wieting et al., 2015, 2016; Wieting and Gimpel, 2018), designing better composition functions (Mitchell and Lapata, 2008; De Boom et al., 2016; Arora et al., 2017; Zhao and Mao, 2017; Rücklé et al., 2018; Zhelezniak et al., 2019b,c; Yang et al., 2019b) and exploring novel similarity measures between word embeddings, in particular those inspired by optimal transport (Kusner et al., 2015; Huang et al., 2016), soft and fuzzy sets (Jimenez et al., 2010, 2015; Zhelezniak et al., 2019b), and statistics (Lev et al., 2015; Nikolentzos et al., 2017; Torki, 2018; Zhelezniak et al., 2019a,c).

Recently, Zhelezniak et al. (2019a,c) advocated for a new statistical perspective on word embeddings where each word embedding itself is viewed as a sample of (e.g. 300) observations from some scalar random variable. They conducted a statistical analysis of several popular pretrained word embeddings and their compositions and established that the ubiquitous cosine similarity is practically equivalent to Pearson correlation. They also demonstrated significant gains in performance when one instead uses non-parametric rank correlation coefficients (Spearman's ρ, Kendall's τ) and cross-covariance operators between reproducing kernel Hilbert spaces (Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005), Centered Kernel Alignment (CKA)) (Cortes et al., 2012; Kornblith et al., 2019).

One prominent alternative to those correlation-based approaches is mutual information (MI), which is of great importance in information theory and statistics. In some sense, mutual information is an excellent candidate for a similarity measure between word embeddings as it can capture arbitrary dependencies between the variables and has a simple and intuitive expression. Unfortunately,

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8361–8371
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

its use in the context of continuous dense word representations has so far been avoided due to the difficulties in estimating MI for continuous random variables (joint and marginal densities are not known in practice).

In this work we make the first steps towards the adoption of MI as a measure of semantic similarity between dense word embeddings. We begin our discussion with how to apply MI for this purpose in principle. Next we carefully summarise the vast literature on estimation of MI for continuous random variables and identify approaches most suitable for our use case. Our chief goal here is to identify the estimators that yield elegant, almost closed-form expressions for the resulting similarity measure as opposed to complicated estimation procedures. Finally, we show that such estimators of mutual information give an excellent signal that correlates very well with human judgements and comfortably rivals existing state-of-the-art unsupervised STS approaches.

2 Background: Statistical Approaches to Word Embeddings

Suppose we are given a word embedding matrix $\mathbf{W} \in \mathbb{R}^{N \times D}$, where N is the vocabulary size and D is the embedding dimension (commonly $D = 300$). Ultimately, the matrix $\mathbf{W}$ is simply a table of some numbers and just like any dataset, it is subject to a statistical analysis. There are essentially two ways we can proceed: we can either choose to view $\mathbf{W}$ as N observations from D random variables or we can instead consider $\mathbf{W}^T$ and view it as D observations from N random variables. The first approach allows us to study 'global' properties of the word embedding space (e.g. via PCA, clustering, etc.) and defines 'global' similarity structures, such as Mahalanobis distance, Fisher kernel (Lev et al., 2015), etc.

In the second approach we study the distribution $P(W_1, W_2, \ldots, W_N)$, where a word embedding $\mathbf{w}_i$ is a sample of D $(= 300)$ observations from some scalar random variable W_i corresponding to the word w_i (Zhelezniak et al., 2019a,c). The 'local' similarity between two words w_i and w_j is then encoded in the dependencies between the corresponding random variables W_i, W_j. Since the distribution $P(W_i, W_j)$ is unknown, we estimate these dependencies based on the sample $\mathbf{w}_i, \mathbf{w}_j$. Certain dependencies can be captured by Pearson, Spearman and Kendall correlation coefficients between word embeddings $\widehat{\rho}(\mathbf{w}_i, \mathbf{w}_j)$, where the choice of the coefficient depends on the statistics of each word embedding model (Zhelezniak et al., 2019a).

Conveniently, correlations can also be used to measure semantic similarity between two sets of words (e.g. phrases and sentences) if one considers the correlations between random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_{l_x})$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{l_y})$, where scalar random variables X_i correspond to the words in the first sentence and Y_j to the words in the second sentence. This, for example, can be done by first pooling (e.g. mean- or max-pooling) random vectors into scalar variables X_{pool} and Y_{pool} and then estimating univariate correlations $\text{corr}(X_{\text{pool}}, Y_{\text{pool}})$ as before. Alternatively, we can measure correlations between random vectors directly using norms of cross-covariance matrices/operators (e.g. the Hilbert-Schmidt independence criterion (Gretton et al., 2005)). Both approaches are known to give excellent results on standard STS benchmarks (Zhelezniak et al., 2019c). A viable alternative to correlations is mutual information (MI), which can detect any kind of dependence between random variables, but which has so far not been explored for this problem.

3 Mutual Information between Dense Word Embeddings

We operate within the previous setting where we consider two sentences $x = x_1 x_2 \ldots x_{l_x}$ and $y = y_1 y_2 \ldots y_{l_y}$. Our goal now is to estimate the mutual information $\mathrm{I}(\mathbf{X}; \mathbf{Y})$ between the corresponding random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_{l_x})$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{l_y})$

$$\mathrm{I}(\mathbf{X}; \mathbf{Y}) = \iint p_{\mathbf{XY}}(x, y) \log \frac{p_{\mathbf{XY}}(x, y)}{p_{\mathbf{X}}(x) p_{\mathbf{Y}}(y)} \mathrm{d}x \mathrm{d}y, \tag{1}$$

where $p_{\mathbf{XY}}(x, y)$ is the joint density of $\mathbf{X}$ and $\mathbf{Y}$ and $p_{\mathbf{X}}(x) = \int_{\mathcal{Y}} p_{\mathbf{XY}}(x, y) \mathrm{d}y$ and $p_{\mathbf{Y}}(y) = \int_{\mathcal{X}} p_{\mathbf{XY}}(x, y) \mathrm{d}x$ are the marginal densities. Unfortunately, these theoretical quantities are not available to us and we must somehow estimate $\widehat{\mathrm{I}}(\mathbf{X}; \mathbf{Y})$ directly from the word embeddings $\widehat{\mathbf{X}} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(l_x)})$ and $\widehat{\mathbf{Y}} = (\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(l_y)})$. Luckily, there is a vast literature on how to estimate mutual information between continuous random variables based on the sample. The first class of methods partitions the supports $\mathcal{X}, \mathcal{Y}$ into a finite number of bins of equal or unequal (adaptive) size and estimates $\widehat{\mathrm{I}}(\mathbf{X}; \mathbf{Y})$

based on discrete counts in each bin (Moddemeijer, 1989; Fraser and Swinney, 1986; Darbellay and Vajda, 1999; Reshef et al., 2011; Ince et al., 2016). While such methods are easy to understand conceptually, they might suffer from the curse of dimensionality (especially when sentences are long) and in some sense violate our desire for an elegant closed-form similarity measure. The next class of methods constructs kernel density estimates (KDE) and then numerically integrates such approximate densities to obtain MI (Moon et al., 1995; Steuer et al., 2002). These methods might require a careful choice of kernels and the bandwidth parameters and also violate our simplicity requirement. The third class of methods that has recently gained popularity in the deep learning community is based on neural-network-based estimation of various bounds on mutual information (e.g. by training a critic to estimate the density ratio in (1)) (Suzuki et al., 2008; Alemi et al., 2017; Belghazi et al., 2018; Hjelm et al., 2019; Poole et al., 2019). Such estimators are usually differentiable and scale well to high dimensions and large sample sizes (Belghazi et al., 2018). However, in our case the sample size (e.g. 300) and dimensionality are not too large (at least for short phrases and sentences), and thus training a separate neural network for a simple similarity computation is hardly justified. This leaves us with the last class of methods that estimates mutual information from the k-nearest neighbour statistics (Kraskov et al., 2004; Ver Steeg and Galstyan, 2013; Ver Steeg, 2014; Ross, 2014; Gao et al., 2015; Gao et al., 2018). These approaches are not without problems (Gao et al., 2015) and inherit the weaknesses of kNN in large dimensions but are very simple to implement. In particular, we focus on the Kraskov–Stögbauer–Grassberger (KSG) estimator (Kraskov et al., 2004) which admits a particularly elegant expression for the resulting similarity measure.

3.1 The KSG Similarity Measure

It can be verified that the mutual information is given by $\mathrm{I}(\mathbf{X};\mathbf{Y}) = \mathrm{H}(\mathbf{X}) + \mathrm{H}(\mathbf{Y}) - \mathrm{H}(\mathbf{X},\mathbf{Y})$, i.e. the difference between the sum of marginal entropies and the joint entropy. Thus, in order to estimate MI, it is sufficient to be able to estimate various entropies in the above equation. In their seminal work, Kozachenko and Leonenko (1987) show how to estimate such differential entropies based on the nearest neighbour statistics. Concretely, these methods approximate the log-density

Algorithm 1 Kraskov–Stögbauer–Grassberger (KSG) Similarity Measure

Require: Word embeddings for the first sentence $\mathbf{X} \in \mathbb{R}^{l_x \times D}$
Require: Word embeddings for the second sentence $\mathbf{Y} \in \mathbb{R}^{l_y \times D}$
Require: The number of nearest neighbours $k < D$ (default $k = 3$)
Ensure: Similarity measure KSG
  $\mathbf{Z} \leftarrow \text{STACK\_ROWS}(\mathbf{X}, \mathbf{Y})$
  $||\mathbf{z}^i - \mathbf{z}^j||_{\mathcal{Z}} \leftarrow \max(||\mathbf{x}^i - \mathbf{x}^j||_{\mathcal{X}}, ||\mathbf{y}^i - \mathbf{y}^j||_{\mathcal{Y}})$
  $i, j = 1, \ldots, D$
  $\# \leftarrow$ set cardinality
  for $\mathbf{z}^d, d = 1, \ldots, D$ **do**
    $\epsilon[d] \leftarrow ||\mathbf{z}^d - \mathbf{z}^{d_k}||, \mathbf{z}^{d_k} = k$-NN of $\mathbf{z}^d$
    $n_x[d] \leftarrow \#\{\mathbf{x}^{d'} : ||\mathbf{x}^d - \mathbf{x}^{d'}||_{\mathcal{X}} < \epsilon[d]\}$
    $n_y[d] \leftarrow \#\{\mathbf{y}^{d'} : ||\mathbf{y}^d - \mathbf{y}^{d'}||_{\mathcal{Y}} < \epsilon[d]\}$
    $d' \in \{1, \ldots D\} \setminus \{d\}$
  end for
  $\psi(x) \leftarrow$ digamma function
  $S \leftarrow \sum_{d=1}^{D} (\psi(n_x[d] + 1) + \psi(n_y[d] + 1))$
  KSG $\leftarrow \psi(D) + \psi(k) - S$

at a point by a uniform density in a e.g. Euclidean or Chebyshev norm ball containing its k-nearest neighbours. Kraskov et al. (2004) modify this idea to construct their famous KSG estimator of mutual information given by

$$\mathrm{KSG}(\mathbf{X};\mathbf{Y}) = \psi(D) + \psi(k) - \sum_{d=1}^{D} \left(\psi(n_x[d] + 1) + \psi(n_y[d] + 1)\right), \quad (2)$$

where D is the embedding dimension, k is the number of nearest neighbours, $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function and $n_x[d], n_y[d]$ are certain nearest neighbour statistics. These statistics are obtained by counting the number of neighbours that fall within less than $\epsilon[d]$ from $\mathbf{x}^d$ and $\mathbf{y}^d$ in the marginal spaces $\mathbf{X}$ and $\mathbf{Y}$ respectively, where $\epsilon[d]$ is the distance from $\mathbf{z}^d = (\mathbf{x}^d, \mathbf{y}^d)$ to its k-nearest neighbour in the joint space $(\mathbf{X}, \mathbf{Y})$. We illustrate how the estimator can be applied to measure similarity between sets of word embeddings in Algorithm 1 and refer the reader to Kraskov et al. (2004) for its full derivation and justification as well as an alternative version.

Similarity	STS	12	13	14	15	16
Popular approaches						
USE (Transf.)		63.8	63.1	66.0	77.1	76.4
BERT Small		50.8	50.4	54.0	62.9	63.8
BERT Large		51.0	47.2	51.8	58.0	62.7
WMD		54.8	47.0	57.7	65.8	63.2
SoftCard		54.8	50.6	58.1	66.5	65.9
DynaMax		61.3	61.7	66.9	76.5	74.7
MeanPool+COS		58.8	58.8	63.4	69.1	68.3
SIF+PCA		58.1	67.2	66.5	73.8	73.0
Correlation-based Approaches						
MaxPool+SPR		61.4	63.8	68.0	75.8	75.9
CKA Gaussian		60.8	64.6	68.0	76.4	73.8
CKA dCorr		60.9	63.4	67.8	76.2	73.4
Mutual Information (KSG)						
KSG $k = 3$		59.9	61.6	67.8	76.7	74.7
KSG $k = 10$		60.4	61.5	68.3	77.0	75.1
MaxPool+KSG 10		59.5	60.2	67.5	75.0	74.1

Table 1: Average Spearman correlation between system and human scores on STS 12–16 tasks. FastText is used for all methods that rely on word embeddings. Similarity measures based on Mutual Information (KSG) perform on par with correlation-based measures and other popular methods from the literature.

4 Experiments

We now explore the empirical performance of the KSG similarity measure on a standard suite of Semantic Textual Similarity (STS) benchmarks (Agirre et al., 2012, 2013, 2014, 2015, 2016) and report Spearman correlation between the system and human scores. Our focus here is on fastText vectors (Bojanowski et al., 2017) trained on Common Crawl (600B tokens), as previous literature suggests that among unsupervised vectors fastText yields the best performance for all tasks and similarity measures (Conneau et al., 2017; Perone et al., 2018; Zhelezniak et al., 2019a,b,c). We defer evaluations and significance analysis on all 24 STS subtasks for other word vectors (word2vec and GloVe) to the Appendix. Our evaluations are run in the SentEval toolkit (Conneau and Kiela, 2018) and our code is available on GitHub[1]. Note that we do not report results on the STS13 SMT subtask as it is no longer publicly available.

[1] https://github.com/babylonhealth/corrsim

Similarity	Time complexity
WMD	$O(m^2 D + m^3 \log m)$
WMD (relaxed)	$O(m^2 D)$
SoftCard	$O(m^2 D)$
DynaMax	$O(m^2 D)$
MaxPool+SPR	$O(mD + D \log D)$
MaxPool+KSG	$O(mD + D^{3/2})$
CKA	$O(mD^2)$
KSG	$O(mD^2)$

Table 2: Computational complexity of some word embedding-based methods, where m is the length of the longer sentence and D is the word embedding dimension.

The number of nearest neighbours for KSG that is known to work well in practice on a variety of datasets is $k = 3$ (Kraskov et al., 2004; Khan et al., 2007). This value seems to strike a good balance between the bias and variance of the estimator. We also run experiments for $k = 10$ to show that KSG is not very sensitive to this hyperparameter, at least in our setting. As an interesting addition, we also run KSG ($k = 10$) for max-pooled scalar random variables (MaxPool+KSG 10). We compare KSG to the following approaches from the literature: Universal Sentence Encoder (Transformer) (Cer et al., 2018), BERT (penultimate layer, mean-pooling) (Devlin et al., 2018), Word Mover's Distance (WMD) (Kusner et al., 2015), soft cardinality (Jimenez et al., 2010, 2015) with cosine similarity and the softness parameter $p = 1$, DynaMax-Jaccard (Zhelezniak et al., 2019b), mean-pooling with cosine similarity (MeanPool+COS) and Smooth Inverse Frequency (SIF) + PCA (Arora et al., 2017). Next we compare KSG with the following top-performing correlations: max-pooling with Spearman correlation (MaxPool+SPR), Centered Kernel Alignment (Gaussian kernel with median estimation for σ^2) and distance correlation (Zhelezniak et al., 2019c). The evaluation results are given in Table 1.

In summary, we can see that similarity measures based on mutual information (KSG) perform on par with top correlation-based measures and other leading methods from the literature. Moreover, KSG between pooled variables (MaxPool) is faster and performs only slightly worse than multivariate KSG.

5 Conclusion

In this work we explored how to apply mutual information (MI) as a semantic similarity measure for continuous dense word embeddings. We have summarised the vast literature on estimating MI for continuous random variables from the sample and singled out a simple and elegant KSG estimator which is based on elementary nearest-neighbour statistics. We showed empirically that this estimator and mutual information in general can be an excellent candidate for a similarity measure between dense word embeddings.

Acknowledgements

We would like to thank Adam Bozson and the four anonymous reviewers for their useful feedback and suggestions.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: 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 (SemEval 2012)*, pages 385–393. Association for Computational Linguistics.

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*, pages 32–43. Association for Computational Linguistics.

Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *ICLR*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *International Conference on Learning Representations*.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmssan, Stockholm Sweden. PMLR.

Yoshua Bengio, Rjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2012. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

G.A. Darbellay and I. Vajda. 1999. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recogn. Lett.*, 80(C):150–156.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Andrew M. Fraser and Harry L. Swinney. 1986. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134–1140.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. 2015. Efficient Estimation of Mutual Information for Strongly Dependent Variables. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 277–286, San Diego, California, USA. PMLR.

W. Gao, S. Oh, and P. Viswanath. 2018. Demystifying fixed k -nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 64(8):5629–5661.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Gao Huang, Chuan Quo, Matt J. Kusner, Yu Sun, Kilian Q. Weinberger, and Fei Sha. 2016. Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4869–4877, USA. Curran Associates Inc.

Robin A.A. Ince, Bruno L. Giordano, Christoph Kayser, Guillaume A. Rousselet, Joachim Gross, and Philippe G. Schyns. 2016. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, 38(3):1541–1573.

Sergio Jimenez, Fabio Gonzalez, and Alexander Gelbukh. 2010. Text comparison using soft cardinality. In *String Processing and Information Retrieval*, pages 297–302, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sergio Jimenez, Fabio A. Gonzalez, and Alexander Gelbukh. 2015. Soft cardinality in semantic text processing: Experience of the SemEval international competitions. *Polibits*, 51:63–72.

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

Shiraj Khan, Sharba Bandyopadhyay, Auroop R. Ganguly, Sunil Saigal, David J. Erickson, Vladimir Protopopescu, and George Ostrouchov. 2007. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2).

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *ICML*.

LF Kozachenko and Nikolai N Leonenko. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Phys. Rev. E*, 69:066138.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 957–966. JMLR.org.

Guy Lev, Benjamin Klein, and Lior Wolf. 2015. In defense of word embedding for generic text representation. In *Natural Language Processing and Information Systems*, pages 35–50, Cham. Springer International Publishing.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244. Association for Computational Linguistics.

R. Moddemeijer. 1989. On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16(3):233–248.

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. 1995. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318–2321.

Giannis Nikolentzos, Polykarpos Meladianos, Francois Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2017. Multivariate Gaussian document representation from word embeddings for text categorization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 450–455, Valencia, Spain. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180, Long Beach, California, USA. PMLR.

Alec Radford. 2018. Improving language understanding by generative pre-training.

D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. 2011. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Brian C. Ross. 2014. Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2):e87357.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *CoRR*, abs/1803.01400.

R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. 2002. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(Suppl 2):S231–S240.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. 2008. Approximating mutual information by maximum likelihood density ratio estimation. In *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*, volume 4 of *Proceedings of Machine Learning Research*, pages 5–20, Antwerp, Belgium. PMLR.

Marwan Torki. 2018. A document descriptor using covariance of word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 527–532, Melbourne, Australia. Association for Computational Linguistics.

Greg Ver Steeg. 2014. Non-parametric entropy estimation toolbox (NPEET).

Greg Ver Steeg and Aram Galstyan. 2013. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 3–12, New York, NY, USA. ACM.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In *International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019a. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Ziyi Yang, Chenguang Zhu, and Weizhu Chen. 2019b. Parameter-free sentence embedding via orthogonal basis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 638–648, Hong Kong, China. Association for Computational Linguistics.

Rui Zhao and Kezhi Mao. 2017. Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems*, pages 1–1.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019a. Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. 2019b. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. In *International Conference on Learning Representations*.

Vitalii Zhelezniak, April Shen, Daniel Busbridge, Aleksandar Savkov, and Nils Hammerla. 2019c. Correlations between word vector sets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 77–87, Hong Kong, China. Association for Computational Linguistics.

Appendix

		GloVe			fastText			word2vec		
		COS	**KSG**	**Δ95% CI**	**COS**	**KSG**	**Δ95% CI**	**COS**	**KSG**	**Δ95% CI**
STS12	MSRpar	**45.17**	**47.79**	[-7.49, 2.09]	44.49	**49.37**	[-9.84, -0.14]	**40.86**	**39.34**	[-2.18, 5.11]
	MSRvid	67.50	**70.95**	[-5.99, -1.04]	73.79	**76.65**	[-4.92, -0.87]	**76.46**	71.87	[2.84, 6.68]
	SMTeuroparl	**58.46**	**55.68**	[-1.67, 7.41]	**62.34**	**58.76**	[-0.10, 7.30]	**49.58**	**47.00**	[-0.44, 5.72]
	surprise.OnWN	61.06	**68.72**	[-10.58, -4.86]	67.50	**70.09**	[-4.73, -0.53]	**67.94**	**68.48**	[-2.13, 1.07]
	surprise.SMTnews	33.92	**45.59**	[-18.19, -5.90]	**45.92**	**47.06**	[-5.27, 2.87]	**42.18**	**44.72**	[-6.51, 1.19]
STS13	FNWN	**36.57**	**45.14**	[-20.76, 2.89]	**39.99**	**48.43**	[-20.07, 2.65]	**40.76**	**46.23**	[-15.47, 4.28]
	headlines	63.12	**70.76**	[-10.50, -5.00]	70.25	**73.15**	[-5.14, -0.93]	**64.09**	**64.88**	[-2.73, 1.04]
	OnWN	**52.57**	**52.58**	[-4.11, 4.01]	**66.26**	62.90	[0.34, 6.71]	**69.84**	62.09	[5.14, 10.80]
STS14	deft-forum	34.72	**48.14**	[-20.66, -6.81]	41.07	**53.65**	[-18.35, -7.35]	**44.57**	**48.82**	[-9.05, 0.34]
	deft-news	**64.56**	**65.85**	[-7.21, 4.80]	**67.46**	**67.37**	[-4.28, 4.52]	**64.00**	**61.22**	[-0.73, 6.95]
	headlines	55.10	**63.42**	[-11.10, -5.83]	61.78	**65.21**	[-5.71, -1.39]	**58.27**	**60.12**	[-3.84, 0.01]
	images	61.26	**74.69**	[-16.90, -10.32]	69.62	**76.88**	[-10.00, -4.69]	**74.60**	**76.01**	[-3.11, 0.30]
	OnWN	**64.35**	**66.84**	[-5.12, 0.07]	**74.48**	**74.15**	[-1.39, 2.07]	**77.85**	73.46	[2.82, 6.14]
	tweet-news	53.83	**72.51**	[-23.66, -14.00]	66.15	**72.54**	[-9.74, -3.28]	65.00	**70.54**	[-8.74, -2.88]
STS15	answers-forums	37.02	**66.53**	[-37.58, -22.43]	56.73	**72.83**	[-22.20, -10.58]	51.26	**63.91**	[-19.15, -6.76]
	answers-students	68.36	**75.16**	[-9.85, -4.32]	**74.15**	**75.34**	[-3.36, 0.71]	**74.55**	**75.03**	[-2.51, 1.28]
	belief	52.76	**72.92**	[-27.61, -13.50]	64.97	**77.97**	[-18.26, -8.35]	65.30	**76.32**	[-16.52, -6.51]
	headlines	66.21	**73.31**	[-9.50, -4.96]	71.85	**75.18**	[-4.98, -1.79]	**67.57**	**68.71**	[-2.74, 0.42]
	images	71.87	**80.35**	[-11.23, -5.96]	77.72	**83.67**	[-8.06, -4.00]	**81.21**	**82.36**	[-2.61, 0.26]
STS16	answer-answer	42.52	**63.01**	[-30.07, -11.93]	49.44	**66.27**	[-25.49, -9.76]	44.41	**60.87**	[-25.21, -9.08]
	headlines	65.88	**73.06**	[-11.77, -3.34]	71.29	**74.83**	[-7.04, -0.66]	**67.90**	**67.38**	[-2.32, 3.16]
	plagiarism	56.10	**80.85**	[-34.89, -16.47]	76.16	**82.10**	[-11.81, -0.80]	**75.90**	**80.58**	[-10.08, 0.14]
	postediting	71.76	**83.57**	[-18.32, -6.74]	78.29	**84.06**	[-10.91, -1.59]	78.94	**83.08**	[-7.66, -1.16]
	question-question	**53.31**	**60.90**	[-16.60, 1.33]	**66.40**	**68.39**	[-8.75, 4.79]	**64.95**	**59.31**	[-1.20, 13.43]

Table 3: **MeanPool+Cosine vs. KGS** ($k = 10$)**:** Spearman correlation between human and system sentence similarity scores. Values in bold indicate the best result on a subtask for a given set of word vectors. The winner is determined based on a 95% BCa confidence interval (Efron, 1987) on the difference in performance between the two systems. When there is no significant difference, both values are in bold.

		GloVe			fastText			word2vec		
		SPR	**KSG**	**Δ95% CI**	**SPR**	**KSG**	**Δ95% CI**	**SPR**	**KSG**	**Δ95% CI**
STS12	MSRpar	41.28	**47.79**	[-9.47, -3.75]	44.79	**49.37**	[-7.36, -2.03]	**36.81**	**39.34**	[-5.10, 0.11]
	MSRvid	**77.32**	70.95	[4.65, 8.48]	**81.76**	76.65	[3.72, 6.74]	**74.14**	71.87	[0.78, 4.00]
	SMTeuroparl	**53.63**	**55.68**	[-4.22, 0.10]	**58.54**	**58.76**	[-2.24, 1.81]	**47.28**	**47.00**	[-1.82, 2.42]
	surprise.OnWN	**68.71**	**68.72**	[-1.29, 1.28]	**71.96**	70.09	[0.60, 3.16]	**67.99**	**68.48**	[-1.80, 0.83]
	surprise.SMTnews	**45.71**	**45.59**	[-3.40, 3.72]	**49.82**	**47.06**	[-0.24, 5.85]	**41.96**	**44.72**	[-5.84, 0.17]
STS13	FNWN	**47.53**	**45.14**	[-5.09, 10.30]	**47.68**	**48.43**	[-10.10, 8.35]	**50.77**	**46.23**	[-5.57, 15.97]
	headlines	**69.45**	**70.76**	[-2.95, 0.34]	**72.23**	**73.15**	[-2.40, 0.52]	**64.29**	**64.88**	[-2.12, 0.89]
	OnWN	**61.98**	52.58	[6.94, 12.29]	**71.43**	62.90	[6.31, 11.24]	**69.68**	62.09	[5.58, 9.93]
STS14	deft-forum	**44.17**	**48.14**	[-8.00, 0.06]	**51.27**	**53.65**	[-5.62, 0.95]	44.23	**48.82**	[-8.32, -1.13]
	deft-news	**66.90**	**65.85**	[-1.48, 3.94]	**65.72**	**67.37**	[-4.69, 1.12]	**59.60**	**61.22**	[-4.45, 1.29]
	headlines	61.58	**63.42**	[-3.50, -0.27]	**64.03**	**65.21**	[-2.81, 0.33]	**58.98**	**60.12**	[-2.76, 0.45]
	images	**75.37**	**74.69**	[-0.72, 2.22]	**77.72**	**76.88**	[-0.47, 2.19]	**75.78**	**76.01**	[-1.57, 1.11]
	OnWN	**72.29**	66.84	[3.94, 7.23]	**77.63**	74.15	[2.23, 4.85]	**77.37**	73.46	[2.68, 5.37]
	tweet-news	70.12	**72.51**	[-4.04, -0.92]	**71.42**	**72.54**	[-2.55, 0.31]	68.09	**70.54**	[-3.86, -1.09]
STS15	answers-forums	**66.02**	**66.53**	[-4.54, 3.53]	69.46	**72.83**	[-7.04, -0.29]	**59.98**	**63.91**	[-8.30, 0.45]
	answers-students	71.34	**75.16**	[-5.49, -2.34]	73.32	**75.34**	[-3.58, -0.54]	**74.48**	**75.03**	[-1.67, 0.54]
	belief	**73.50**	**72.92**	[-2.25, 3.71]	**77.69**	**77.97**	[-2.83, 2.48]	**73.53**	**76.32**	[-5.69, 0.28]
	headlines	71.77	**73.31**	[-2.85, -0.32]	**74.17**	**75.18**	[-2.20, 0.10]	**67.87**	**68.71**	[-2.13, 0.43]
	images	**81.94**	80.35	[0.33, 2.88]	**84.49**	**83.67**	[-0.12, 1.78]	**82.60**	**82.36**	[-0.68, 1.21]
STS16	answer-answer	**61.30**	**63.01**	[-5.73, 2.46]	**65.98**	**66.27**	[-4.08, 3.97]	**59.09**	**60.87**	[-5.43, 1.76]
	headlines	70.03	**73.06**	[-5.17, -1.24]	72.96	**74.83**	[-3.96, -0.00]	**67.87**	**67.38**	[-1.12, 2.31]
	plagiarism	77.72	**80.85**	[-5.93, -0.98]	**83.75**	**82.10**	[-0.09, 4.08]	**80.28**	**80.58**	[-2.44, 1.55]
	postediting	81.45	**83.57**	[-3.69, -0.77]	**82.85**	**84.06**	[-2.94, 0.35]	80.06	**83.08**	[-4.96, -1.37]
	question-question	**66.80**	60.90	[1.23, 11.53]	**74.03**	68.39	[2.14, 10.06]	**65.87**	59.31	[1.37, 13.10]

Table 4: **MaxPool+Spearman vs. KGS** ($k = 10$)**:** Spearman correlation between human and system sentence similarity scores. Values in bold indicate the best result on a subtask for a given set of word vectors. The winner is determined based on a 95% BCa confidence interval (Efron, 1987) on the difference in performance between the two systems. When there is no significant difference, both values are in bold.

		GloVe			fastText			word2vec		
		CKA	**KSG**	**Δ95% CI**	**CKA**	**KSG**	**Δ95% CI**	**CKA**	**KSG**	**Δ95% CI**
STS12	MSRpar	42.65	**47.79**	[-7.97, -2.60]	45.12	**49.37**	[-6.64, -2.18]	36.00	**39.34**	[-5.92, -1.01]
	MSRvid	**76.93**	70.95	[4.47, 7.77]	**83.78**	76.65	[5.66, 8.89]	**79.64**	71.87	[6.23, 9.62]
	SMTeuroparl	**57.62**	55.68	[0.33, 3.81]	**58.74**	**58.76**	[-2.18, 2.02]	**46.80**	**47.00**	[-1.84, 1.46]
	surprise.OnWN	66.21	**68.72**	[-3.82, -1.16]	68.10	**70.09**	[-3.31, -0.70]	66.36	**68.48**	[-3.42, -0.91]
	surprise.SMTnews	**46.94**	**45.59**	[-1.19, 3.88]	**48.45**	**47.06**	[-1.62, 4.46]	**44.12**	**44.72**	[-3.70, 2.09]
STS13	FNWN	**37.98**	**45.14**	[-15.29, 1.05]	**48.85**	**48.43**	[-8.13, 9.24]	**42.02**	**46.23**	[-12.02, 3.58]
	headlines	**70.34**	**70.76**	[-1.94, 1.00]	71.65	**73.15**	[-2.96, -0.17]	63.02	**64.88**	[-3.28, -0.47]
	OnWN	**61.35**	52.58	[6.67, 11.17]	**73.46**	62.90	[8.19, 13.37]	**71.23**	62.09	[7.11, 11.57]
STS14	deft-forum	**50.80**	**48.14**	[-0.62, 6.11]	**53.67**	**53.65**	[-3.66, 3.57]	**51.43**	**48.82**	[-0.94, 6.17]
	deft-news	**67.78**	**65.85**	[-1.13, 5.22]	**67.18**	**67.37**	[-3.07, 2.76]	**61.48**	**61.22**	[-2.45, 3.31]
	headlines	61.51	**63.42**	[-3.41, -0.51]	63.47	**65.21**	[-3.22, -0.35]	58.31	**60.12**	[-3.36, -0.37]
	images	**74.08**	**74.69**	[-2.02, 0.79]	**77.50**	**76.88**	[-0.49, 1.84]	**76.44**	**76.01**	[-0.65, 1.54]
	OnWN	**72.14**	66.84	[4.00, 6.77]	**79.28**	74.15	[3.76, 6.63]	**78.45**	73.46	[3.79, 6.46]
	tweet-news	67.22	**72.51**	[-7.62, -3.32]	66.81	**72.54**	[-8.07, -3.85]	65.75	**70.54**	[-6.73, -3.15]
STS15	answers-forums	**64.46**	**66.53**	[-4.87, 0.52]	**73.62**	**72.83**	[-1.30, 2.99]	**62.50**	**63.91**	[-4.09, 1.24]
	answers-students	73.23	**75.16**	[-3.86, -0.21]	72.11	**75.34**	[-5.03, -1.69]	**73.90**	**75.03**	[-2.58, 0.10]
	belief	**71.67**	**72.92**	[-4.59, 2.19]	**76.50**	**77.97**	[-4.26, 1.14]	**74.04**	**76.32**	[-5.15, 0.17]
	headlines	**73.10**	**73.31**	[-1.36, 0.90]	**74.60**	**75.18**	[-1.76, 0.53]	**67.90**	**68.71**	[-2.01, 0.41]
	images	**81.48**	**80.35**	[-0.16, 2.44]	**85.04**	83.67	[0.46, 2.34]	**83.75**	82.36	[0.49, 2.38]
STS16	answer-answer	55.29	**63.01**	[-13.91, -2.04]	61.19	**66.27**	[-10.28, -0.36]	52.34	**60.87**	[-14.13, -3.81]
	headlines	70.79	**73.06**	[-4.30, -0.38]	72.35	**74.83**	[-4.44, -0.59]	65.16	**67.38**	[-4.27, -0.34]
	plagiarism	**79.90**	**80.85**	[-4.24, 1.71]	**80.19**	**82.10**	[-4.65, 0.16]	**80.53**	**80.58**	[-1.80, 1.86]
	postediting	81.37	**83.57**	[-4.92, -0.14]	81.96	**84.06**	[-4.12, -0.49]	80.85	**83.08**	[-4.24, -0.42]
	question-question	**72.46**	60.90	[6.29, 18.67]	**73.32**	68.39	[1.05, 10.39]	**70.08**	59.31	[5.74, 17.83]

Table 5: **CKA (Gaussian) vs. KGS** ($k = 10$)**:** Spearman correlation between human and system sentence similarity scores. Values in bold indicate the best result on a subtask for a given set of word vectors. The winner is determined based on a 95% BCa confidence interval (Efron, 1987) on the difference in performance between the two systems. When there is no significant difference, both values are in bold.