Hypothesis Testing based Intrinsic Evaluation of Word Embeddings

Nishant Gurnani
Department of Mathematics
University of California San Diego
ndgurnan@ucsd.edu

Abstract

We introduce the cross-match test - an exact, distribution free, high-dimensional hypothesis test as an intrinsic evaluation metric for word embeddings. We show that cross-match is an effective means of measuring distributional similarity between different vector representations and of evaluating the statistical significance of different vector embedding models. Additionally, we find that cross-match can be used to provide a quantitative measure of linguistic similarity for selecting bridge languages for machine translation. We demonstrate that the results of the hypothesis test align with our expectations and note that the framework of two sample hypothesis testing is not limited to word embeddings and can be extended to all vector representations.

1 Introduction

Word embeddings obtained via specialized models (Brown et al., 1992; Pennington et al., 2014; Mikolov et al., 2013a) or neural networks (Bengio et al., 2003) have been successfully used to address various natural language processing tasks (Vaswani et al., 2013; Soricut and Och, 2015). These embeddings provide a nuanced representation of words that can capture various syntactic and semantic properties of natural language (Mikolov et al., 2013b). Despite their effectiveness in downstream applications, embeddings have limited practical value as standalone items. Consequently, an intrinsic evaluation metric must provide insight on the downstream task the embeddings are designed for. In this work, we use Cross-match (Rosenbaum, 2005) - an exact, distribution free, high-dimensional hypothesis test to propose a novel approach for intrinsic evaluation of word embeddings, one that provides insight on tasks that depend on linguistic similarity.

Evaluating general purpose vector representations is difficult. They are trained using simple objectives and applied to a variety of downstream tasks, thus making no single extrinsic evaluation definitive. Often, due to computational constraints, direct downstream evaluations are also impractical. In the case of word embeddings, these constraints have led to the development of dedicated evaluation tasks like similarity and analogy (Rohde et al., 2006; Levy et al., 2015) which are not directly related to training objectives or to downstream tasks. Despite their ease of interpretability, Faruqui et al. (2016) have shown that these tasks do not correlate well with downstream performance. In related work, Tsvetkov et al. (2016) propose an evaluation measure QVEC-CCA that is shown to correlate well with downstream semantic tasks where the objective is to quantify the linguistic content of word embeddings by maximizing the correlation with a manually annotated linguistic resource.

In this work, we use the Cross-match hypothesis test (Rosenbaum, 2005) to measure distributional similarity between different word vector representations. Cross-match is an adjacency based test traditionally used in clinical settings where the goal is to assess no treatment effect on a high-dimensional outcome in a randomized experiment. In our setting, we assume there exists some unknown distribution W from which our constructed word embeddings $\{\mathbf{w_1}, \ldots, \mathbf{w_n}\}$ are "sampled" from. Given two sets of word embeddings, cross-match tests whether the underlying distribution from which the embeddings were "sampled" are identical or not. The test uses optimal non-bipartite matching to pair vectors from both sets of embeddings based on distance (e.g.

Proceedings of the 2nd Workshop on Evaluating Vector-Space Representations for NLP, pages 16–20,
Copenhagen, Denmark, September 7–11, 2017.

a vector will be paired with it's nearest neighbor based on some distance metric). The cross-match test statistic C is the number of times that a vector from one set is paired with a vector from another. The null hypothesis assumes that the vectors were sampled from the same distribution and rejects for small values of C. Thus, a large number of cross-matches between two sets of word embeddings suggests that they are from the same embedding distribution.

Using cross-match, we propose two illustrative examples of intrinsic evaluation. First, we use pretrained word vectors (trained on Wikipedia using the skip-gram model in Bojanowski et al. (2016)) from Facebook's fastText library for several languages to calculate the cross-match statistic for several language pairs. We hypothesize that for linguistically similar languages, a larger statistic will be observed. Secondly, we use cross-match to assess the statistical significant of word embedding models. We consider several well known models trained on the same corpus and use cross-match to assess whether the respective word vector representations are statistically significantly different. We hypothesize that the number of cross-matches between two different embedding models is small, thus suggesting that they capture fundamentally different linguistic aspects of the corpus.

This paper is organized as follows: Section 2 introduces the cross-match test in detail. Experiments on embedding similarity and evaluation are described in Section 3. We discuss extensions and conclude in Section 4.

2 Cross-Match Test

The cross-match test (Rosenbaum, 2005) is a nonparametric goodness-of-fit test in arbitrary dimensions. It is an exact, distribution-free, two-sample hypothesis test that measures whether two distributions are equal or not. Formally, given two independent samples $w_1, \ldots, w_n \sim W$ and $v_1, \ldots, v_m \sim V$, cross-match tests the null hypothesis $H_0 : W = V$ versus the alternative hypothesis $H_1 : W \neq V$. The test has been traditionally used in clinical settings, where the goal is to assess no treatment effect on a high-dimensional outcome between control and treated subjects in a randomized experiment (Heller et al., 2010). In the case of word embeddings, the goal is to test whether two sets of word embedding vectors have been "sampled" from the same distribution.

2.1 Definition of the Cross-Match Statistic

Let W, V denote two word embedding distributions (distributions of word embedding vectors over a corpus), suppose we obtain two sets of word vectors $\{\mathbf{w_1}, \ldots, \mathbf{w_n}\} \sim W$ and $\{\mathbf{v_1}, \ldots, \mathbf{v_m}\} \sim V$. Assign the group labels 0 and 1 to indicate which sample the vectors are from such that the data are organized as follows: $\{(0, \mathbf{w_1}), \ldots, (0, \mathbf{w_n})\}$ and $\{(1, \mathbf{v_1}), \ldots (1, \mathbf{v_m})\}$.

The cross-match statistic C, is a function of the word vectors $D = \{\mathbf{w_1}, \ldots, \mathbf{w_n}, \mathbf{v_1}, \ldots, \mathbf{v_n}\}$ and the group labels $G = \{0, \ldots, 0, 1, \ldots, 1\}$. If $H_0 : W = V$ is true, then all the word vectors are i.i.d. "sampled" from W and the group labels are meaningless. It's as if the 0's and 1's were randomly assigned.

The cross-match test is performed as follows. For notational convenience ignore the group labels and treat the data as one sample $\{\mathbf{z_1}, \ldots, \mathbf{z_{n+m}}\}$ of size $n+m = N$ (assume for simplicity that N is even). We define a $N \times N$ symmetric distance matrix, with row k and column l giving the distance (any distance metric can be used) between $\mathbf{z_k}$ and $\mathbf{z_l}$. Compute the optimal non-bipartite matching of the $\mathbf{z}'\mathbf{s}$ (match the vectors into non-overlapping pairs) that minimizes the total distances between the points in each pair.

Formally, we find a permutation $\hat{\sigma}$ of $\{1, \ldots, N\}$ that minimizes

$$Match(\sigma) = \sum_{i=1}^{N} d(Z_i, Z_{\sigma(i)})$$

where $i \neq \sigma(i)$ and d is our chosen distance measure. The cross-match statistic C, is defined as the number of pairs that have group labels (0,1) or (1,0), the test rejects for small values of C.

If there is an odd number of word embedding vectors, then a psuedo-vector is added to the distance matrix at zero distance from everyone else. $\frac{N}{2}$ pairs are formed as before, and the pair containing the psuedo-vector is discarded (thus the least matchable word vector is discarded).

2.2 Null Distribution of the Cross-Match Statistic

One advantage of the cross-match test is that we can compute the exact distribution of the statistic C under the null hypothesis H_0. Given $\frac{N}{2}$ paired vectors, let c_0 denote the observed number of the

pairs with group labels (0,0), let c_1 denote the observed number of pairs with group labels (0,1) or (1,0) (this is our observed cross-match statistic) and finally let c_2 denote the observed number of pairs with group labels (1,1). The null distribution of C in closed form is:

$$f(c_1) = P(C = c_1) = \frac{2^{c_1} n!}{\binom{N}{n} c_0! c_1! c_2!}$$

where $\frac{N}{2} = c_0 + c_1 + c_2$. Having the null distribution in closed form also allows us to compute the exact p-value for our observed cross-match statistic. The resulting p-value is equal to $F(c_1)$ where

$$F(c_1) = P(C \leq c_1) = \sum_{c_1'=0}^{c_1} f(c_1')$$

A low p-value would suggests that we have evidence to reject the null hypothesis (at a given level of significance) that the word embedding vectors were "sampled" from the same distribution.

3 Experiments

In the following experiments, we demonstrate two different illustrative examples of the cross-match test. Our objective is to show the effectiveness of cross-match as a general tool for intrinsic evaluation of word embedding vectors.

3.1 Embedding Similarity

A bridge language (also referred to as a pivot language), is an artificial or natural language used as an intermediary for translation between two different languages. In machine translation, a bridge language is useful in low-resource situations where a good parallel corpora is not available for the target language. In such cases, a resource rich, linguistically similar language is used as a proxy in order to perform the required NLP task. For example in Tsvetkov and Dyer (2015) the authors use Arabic, Italian and French as bridge languages to perform Swahili-English, Maltese-English and Romanian-English translations respectively.

Assessing whether languages are linguistically similar is a reasonably difficult task and depends on the notion of similarity one uses (lexical, morphological etc.) In this experiment, we use cross-match to provide a quantitative measure to assess linguistic similarity between languages.

We use pre-trained word vectors (trained on Wikipedia using the skip-gram model in Bojanowski et al. (2016)) from Facebook's fastText library for several languages and calculate the cross-match statistic for several language pairs. Specifically, we randomly select 100,000 word vectors for each language (with the exception of Maltese and Swahili which have only 26,000 and 52,000 vectors respectively). Then for each language pair, we randomly sample 200 vectors and calculate the number of cross-matches between them using R's crossmatch package (`https://github.com/cran/crossmatch`). We repeat this 500 times for each language pair and report the average cross-match statistic.

Language Pair	Cross-Match
English-French	23.76
English-Italian	25.04
English-Spanish	23.36
English-Portuguese	18.44
English-Arabic	19.34
English-Maltese	7.84
English-Romanian	16.56
English-Swahili	17

Table 1: fastText vectors cross-match statistics for English-pair languages

Language Pair	Cross-Match
Maltese-English	7.84
Maltese-French	7.28
Maltese-Italian	9.20
Maltese-Spanish	6.76
Maltese-Portuguese	4.84
Maltese-Arabic	6.68
Maltese-Romanian	6.96
Maltese-Swahili	4.44

Table 2: fastText vectors cross-match statistics for Maltese-pair languages

Tables 1 and 2 present the results of calculating the average number of cross-matches between several English-pair and Maltese-pair languages. We note that with a sample of 400 vectors (200 from each language) the maximum possible number of cross-matches is 200. Given that are our reported statistics are considerably lower than 200 we can safely conclude that the distributions from which the word embedding vec-

tors were generated are different for different languages. In table 1 we note that the number of cross-matches between English and other romance languages (French, Italian, Spanish, Portuguese, Romanian) is noticeably higher than that between English and non-romance languages (Arabic, Maltese, Swahili). This corresponds with our notions of linguistic similarity between the languages, we certainly expect English to be more "similar" to French than to Maltese. We also note that in table 2, the Maltese-Italian pair has the highest cross-match statistic, thus supporting the choice of Italian as a bridge language for Maltese.

3.2 Embedding Evaluation

In this experiment, we use cross-match to assess the statistical significance of word embedding models. Despite the popularity of various different embedding models (Mikolov et al., 2013a,b; Pennington et al., 2014) it is not always clear whether one model represents a statistically significant improvement to other existing models (it maybe that all of them capture largely similar features of the text).

We consider four popular word embedding models: word2vec Skip-gram, word2vec CBOW, Glove and fastText all trained on the same English wikipedia corpus. Once again we take samples of size 200 from each method, caluclate the p-value between two pairs of methods using cross-match and then report the average p-value across 500 repeated iterations.

	Skip	CBOW	Glove	FastText
Skip	-	4.93e-26	2.39e-27	1.66e-23
CBOW	4.93e-26	-	9.42e-25	2.71e-22
Glove	2.39e-27	9.42e-25	-	1.13e-23
fastText	1.66e-23	2.71e-22	1.13e-23	-

Table 3: p-values calculated using Cross-match

The results in 3 show low p-values across all pairs of word embedding methods thus suggesting that they all seem to capture different aspects of the corpus they are modeling. In other words, using cross-match we have evidence to reject the null hypothesis that the vectors derived from any pair of models come from the same word embedding distribution.

Lastly, we note that there are at present some computational constraints in performing the cross-match test. There exists a bottleneck in the calculation of the optimal non-bipartite matching and this makes performing the test for larger sample sizes currently intractable. However, we feel confident that this software issue can be easily overcome by writing custom routines (as opposed to using existing open-source code) and parallelizing the problem. As a result of our limited sample size, we not that it is possible that the power of our hypothesis test is low and thus we may be making type I errors (falsely rejecting the null). Nonetheless our initial results seem promising and are in line with our expectations.

4 Conclusion

In this work we introduced the cross-match test, an exact, distribution free, high-dimensional hypothesis test as an intrinsic evaluation metric for word embeddings. We were able to demonstrate on two illustrative examples that the test performs reasonably in line with our expectations and can potentially be a useful tool in assessing bridge languages for machine translation. Despite the initially promising results, much further work remains to be done in order to confirm the efficacy of cross-match in the context of word embeddings.

We posit that our main contribution is the introduction of the hypothesis testing framework as a method for intrinsic evaluation of vector representations. We observe that there is nothing notable about word embeddings or the cross-match test and our experiments could be extended for other vector representations (sentence, phrase etc.) using other modern two-sample hypothesis tests such as the popular maximum mean discrepancy (Gretton et al., 2012). Given the rich literature on hypothesis testing in statistics, there is certainly much to be explored here.

For future work we aim to focus solely on the problem of bridge languages in machine translation. Our objective is to conduct a larger scale study that is able to definitively show a strong correlation between the results of a hypothesis test on word embedding vectors, and their subsequent performance on the downstream machine translation task.

Acknowledgments

We thank Ndapa Nakashole for several useful discussions helping formulate the problem and the anonymous reviewers for their feedback.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.* 18(4):467–479.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *CoRR* abs/1605.02276.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.* 13:723–773.

Ruth Heller, Shane T. Jensen, Paul R. Rosenbaum, and Dylan S. Small. 2010. Sensitivity analysis for the cross-match test, with applications in genomics. *Journal of the American Statistical Association* 105(491):1005–1013.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurence. *COMMUNICATIONS OF THE ACM* 8:627–633.

Paul R. Rosenbaum. 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(4):515–530.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1627–1637.

Yulia Tsvetkov and Chris Dyer. 2015. Cross-lingual bridges with models of lexical borrowing. *J. Artif. Intell. Res. (JAIR)* 55:63–93.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based intrinsic evaluation of word vector representations. *CoRR* abs/1606.06710.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *EMNLP*. ACL, pages 1387–1392.