Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks

Jeremy Auguste[1], **Arnaud Rey**[2], **Benoit Favre**[1]
Aix-Marseille Université, CNRS
[1]LIF UMR 7279, [2]LPC UMR 7290
13000 Marseille, France
{firstname.lastname}@univ-amu.fr

Abstract

This work presents a framework for word similarity evaluation grounded on cognitive sciences experimental data. Word pair similarities are compared to reaction times of subjects in large scale lexical decision and naming tasks under semantic priming. Results show that GloVe embeddings lead to significantly higher correlation with experimental measurements than other controlled and off-the-shelf embeddings, and that the choice of a training corpus is less important than that of the algorithm. Comparison of rankings with other datasets shows that the cognitive phenomenon covers more aspects than simply word relatedness or similarity.

1 Introduction

Word representations have attracted a lot of interest in the community and led to very useful applications in a range of domains of natural language processing. Such representations are typically evaluated intrinsically on word similarity tasks and extrinsically on their impact on NLP systems performance (Schnabel et al., 2015; Lai et al., 2016; Ghannay et al., 2016).

A recent trend towards building more general representations has looked at how similarities in the representation space can predict the outcome of cognitive experiments, such as human reaction time in semantic priming experiments (Ettinger and Linzen, 2016) or relying on eye tracking and brain imaging data (Søgaard, 2016; Ruan et al., 2016). The idea is that ground truth from unconscious phenomena might be less prone to subjective factors of more traditional word similarity and relatedness datasets.

In this paper, we describe an evaluation framework based on comparing word embedding similarity against reaction times from the Semantic Priming Project (Hutchison et al., 2013). A set of word embeddings is evaluated by computing its Spearman rank correlation with average reaction times obtained by submitting a set of subjects to a prime (one word from the pair) and then perform one of two tasks: lexical decision (decide whether the second word is an existing word or not), and naming (read aloud the second word).

Extending the ideas developed in (Ettinger and Linzen, 2016), this paper describes the following contributions:

- we create and distribute a package[1] for word embedding evaluation based on the SPP primed reaction time data;
- in order to calibrate results from that evaluation framework, we look at the effect of training corpus on a set of word embeddings;
- we also look at the correlation between SPP reaction times and subjective similarity and relatedness ratings from existing datasets.

2 Related Work

Intrinsic and extrinsic approaches have been proposed for word embedding evaluation. The former typically consist in collecting human judgment of word similarity on a range of word pairs, and computing the rank correlation of their averaged value with the cosine similarity between the embeddings of the words in the pair. Word analogy is also evaluated but it has been proven to be equivalent to a linear combination between cosine similarities (Levy et al., 2014). In this work, we focus on

[1]Available at https://github.com/JomnTAL/spp-wordsim

Proceedings of the 2nd Workshop on Evaluating Vector-Space Representations for NLP, pages 21–26,
Copenhagen, Denmark, September 7–11, 2017.

intrinsic evaluation, and therefore do not detail efforts for extrinsic evaluation of word embeddings. More pointers can be found in (Schnabel et al., 2015; Lai et al., 2016; Ghannay et al., 2016).

A number of datasets can be used to evaluate word similarity based on human judgment. Older datasets, such as RG (Rubenstein and Goodenough, 1965) and MC (Miller and Charles, 1991) shall not be used because differences between correlations are not significant due to their small size (Faruqui et al., 2016). While early datasets contained judgments collected in-house, such as WS-353 (Finkelstein et al., 2001; Agirre et al., 2009), most recent ones are created through crowd sourcing, with 10 to 20 annotations per word-pair, filtered for outliers: MTurk-287 (Radinsky et al., 2011), MTurk-771 (Halawi et al., 2012), MEN (Bruni et al., 2012), RW (Luong et al., 2013). Another trend is to address different syntactic categories than noun, such as YP-130 (Yang and Powers, 2006) and Verb (Baker et al., 2014) which focus on verbs. While past studies did not differentiate relatedness from semantic similarity, SimLex (Hill et al., 2016) and SimVerb (Gerz et al., 2016) explicitly promote the later. Embeddings can also be compared to lexical resources (Tsvetkov et al., 2015), but it is hard to balance the contribution of each linguistic phenomenon. Existing similarity ratings are subjective because they are the result of a conscious process while it seems desirable to directly evaluate embeddings against basic processes that support language in the brain.

In the cognitive sciences community, there has been efforts to explain how word associations are formed by introducing word embeddings in models. Pereira et al. (2016) look at how off-the-shelf word embeddings predict free associations (given a prime, say the first word that comes to your mind), a conscious process while we are interested in an unconscious process. The work by Hollis and Westbury (2016), among other experiments, compares word embedding principal components to unprimed reaction times from the English Lexicon Project (ELP) and British Lexicon Project (BLP). Both papers are interested in explaining cognitive behavior. There have also been efforts to evaluate word embeddings against fMRI recordings with the idea that embeddings can explain part of the neural activity (Søgaard, 2016).

The work by Ettinger and Linzen (2016) is particularly relevant to our study in that the authors also propose to evaluate embeddings against reaction time data from the SPP dataset. Our work differs in that we provide an evaluation setup which can be reused by other researchers, we propose a different evaluation metric based on rank correlation, and we perform an analysis in regard to different parameters, in particular the corpus used to train embeddings.

3 Semantic priming

There is an extensive psychological literature concerning the nature of semantic representations and the influence of semantic or associative context on word processing (McNamara, 2005). In this domain, the semantic priming paradigm is one of the most popular experimental tool to study these cognitive processes. In this task, participants are presented with a prime (stimulus) word (e.g., cat) immediately followed by either a related (e.g., dog) or an unrelated (e.g., truck) target word. A speeded response is expected on the target word (e.g., a lexical-decision, i.e., is it a word or not?) and a response time is recorded. Semantic priming refers to the finding that people respond faster to target words preceded by related, relative to unrelated, primes. This behavioral index therefore provides information about the influence of a semantic context (i.e., the prime word) on the processing of the target word and is suitable for the development of theories of semantic memory.

Due to the increasing precision of computational models of word processing, researchers are now testing models by using large-scale databases providing experimental data at the item level. The Semantic Priming Project (SPP) (Hutchison et al., 2013) is one of these recently collected databases. It provides response times from 768 participants in speeded naming (NT) and lexical decision (LDT) tasks for 1,661 target words following related and unrelated primes. The naming task consists in reading aloud the target, while the lexical decision task consists in pressing one of two buttons to specify if the target is a valid word or not. Aside from the relatedness between the prime and the target, the item data is also available for stimulus onset asynchronies (SOA) between the prime and target items of 200 and 1,200 ms. Stimulus onset asynchrony is the time between the end of showing the first (stimulus) word, and the target word in the reaction time experiments. Se-

mantic priming at shorter SOAs (e.g., under 300 ms) is thought to reflect automatic priming mechanisms, whereas priming at longer SOAs (e.g., over 300 ms) presumably reflects additional intentional strategies (Hutchison et al., 2001).

Reaction times observed in semantic priming experiments can be explained by a range of linguistic phenomena, such as relatedness, semantic similarity, syntactic traits, or morphology. Unlike Hill et al. (2016) who focus on one phenomenon, we assume that word representations should convey the full mixture of explaining factors observed in unplanned human behavior. Therefore we evaluate embeddings by computing the correlation between the cosine similarity of pairs of words and reaction times (RT). Shorter RT indicate more priming effect, leading to negative correlations. In addition, non linguistic factors such as frequency are known to influence RT measurements, so it is not expected that word embeddings explain the whole variance of the experiment.

The SPP data is significantly larger than word similarity datasets and consists of 6,637 word pairs. The RT for each pair is an average over the performance of 30 subjects. In addition to reaction times, the dataset contains demographics, proficiency and attention tests results.

4 Evaluation framework

Embeddings are evaluated by computing the cosine similarity between word pairs from the SPP project, and look at their Spearman rank correlation with the RT data. The results are given in term of negative correlation. Significance of the difference between correlations is calculated with the Steiger test (Steiger, 1980).

In this evaluation framework, the word pairs for the Lexical Decision Task (LDT) and Naming Task (NT) are split according to two partitions. The first one (P1), used here, consists of a development set of 1,328 pairs and a test set of 5,309 pairs. Parameters of the proposed approaches can be tuned on the development set and performance must be reported on the test set. Another partition (P2) is made available to also include data for training algorithms. The Train set consists of 3,981 pairs, the Dev has 1,328 pairs and the Test has 1,328 pairs. These partitions were obtained by using 10 folds which are also available. We didn't create them with a particular goal in mind but we made sure that the mean word frequency and the standard deviation (SD) of each fold was close to the mean word frequency and SD of the full dataset. By standardizing the data splits, we ensure that results presented in future work will be comparable.

Experiments In a first set of experiments, we benchmark a range of embeddings on the SPP data. Four conditions are considered: LDT-200, LDT-1200, NT-200, NT-1200 (for lexical decision task, and naming task, both with an onset of 200 ms and 1,200 ms). We compare two categories of embeddings: a controlled setting for which algorithms are trained on the same dataset, and a selection of pretrained embeddings available to the community. Due to space constraints, the pretrained embeddings considered[2] are limited to: W2V Skip-gram (Mikolov et al., 2013), GloVe (Pennington et al., 2014), Multilingual (Faruqui and Dyer, 2014), Dependency-based (Levy and Goldberg, 2014), and FastText (Bojanowski et al., 2016).

For the embeddings with controlled settings, we used two algorithms: W2V Skip-grams and GloVe. We used three different corpora to train these embeddings: Wikipedia 2013 (Wiki), Gigaword[3] (GW) and OpenSubtitles 2016 (OS). We used a centered window of size 10 and generated vectors with 100 dimensions for all 6 models.

From the experiment detailed in Table 1, it appears that GloVe leads to significantly larger negative correlation compared to other approaches, both on the controlled and pretrained settings. On the controlled setting, we notice that the choice of the corpora doesn't significantly affect the correlation. Even if the correlation seems to be higher with the NT-200 and NT-1200 datasets when using the OpenSubtitles corpus, the Steiger test shows that the difference in correlation when using the other corpora is not significant in most cases. However, the algorithms used do have a significant impact on the correlations. In the future, it would be interesting to look at the impact of various other settings such as the size and position of the window, or the dimension of the word vectors.

It can also noted that the correlations on the lexical decision tasks are higher than on the naming tasks which supports the idea that lexical decision

[2]URLs and descriptions available at https://github.com/JomnTAL/spp-wordsim.

[3]LDC2012T21

	Off-the-shelf embeddings					Embeddings with controlled settings					
	GloVe	W2V	Multilingual	Dependency	FastText	GloVe OS	GloVe Wiki	GloVe GW	W2V OS	W2V Wiki	W2V GW
LDT-200	**25.02**	15.35*	13.88*	5.48*	14.48*	23.61	23.06	**23.91**	17.77*	16.75*	16.86*
LDT-1200	**18.91**	11.21*	11.19*	3.76*	10.75*	17.65	17.88	**18.17**	12.86*	11.64*	11.35*
NT-200	**15.43**	5.70*	6.43*	-1.56*	3.73*	**15.37**	12.52*	13.67†	8.27*	6.30*	7.83*
NT-1200	**12.57**	8.86*	7.58*	4.68*	8.30*	**12.14**	10.36†	11.78	9.42*	9.23*	9.30*

Table 1: Spearman's correlation[4] between the test SPP datasets and word embedding models. Highest results are in bold. Significativity of the figures compared to the best results according to Steiger test is indicated by $^*(pval < 0.01)$ and $^\dagger(pval \in [0.01, 0.05[)$.

	Nb Pairs	LDT-200	LDT-1200	NT-200	NT-1200
WS-353-ALL	16	0.40	-0.30	0.26	-0.51
MTurk-771	26	-0.08	0.23	-0.23	-0.28
MEN-TR-3k	71	0.21	-0.20	0.04	-0.02
SimLex-999	101	-0.03	0.06	0.06	0.02

Table 2: Spearman's correlation between reaction times from the SPP datasets and relationship scores from other datasets (with more than 16 overlapping pairs).

seems to be a task less subject to variability from production of the response (pressing a button vs saying a word). Better correlations at an onset of 200 ms can be explained by the fact that subjects are allowed more time to build an intent, leading to more factors being involved. This also probably means that most word embedding models are better at capturing automatic priming mechanisms.

The second experiment detailed in Table 2 looks at the characteristics of the SPP data in regard to other available datasets. We calculated the correlation between the reaction times in the SPP datasets and the relationship scores in a set of existing datasets, using the pairs of words that are available in the two compared datasets (different pairs are compared for each dataset). We only show the results for datasets that have 16 or more pairs in common with the SPP dataset. The SimVerb dataset had 208 pairs in common but since the words used were only verbs whereas in the SPP dataset the part-of-speech wasn't specified, we couldn't really compare the two datasets. It can be observed that the correlations are low which probably means that evaluating on the SPP data could outline different phenomena than what is already covered by relatedness and similarity oriented datasets. Additional work has to be done to fully understand what factors are taken into account by the SPP data.

[4]For readability, correlations have been multiplied by -100

5 Discussion

It is not clear what cognitive processes lead to mental representations of words in the brain, and it is not clear how these processes relate to linguistic theories. However, reaction time in the context of semantic priming seems to be a good proxy for modeling word embeddings after cognitive processes. The proposed framework addresses some of the problems with word embedding evaluation exposed in (Faruqui et al., 2016)[5]: (2.1) subjectivity is addressed by looking at unconscious phenomena, (2.2) the lexical decision and naming tasks are very general but they are affected by other cognitive pipelines such as vision, (2.3) we provide standardized splits, and (2.5) the size of the dataset allows for significant differences between algorithms. However, we do not address the problem of low correlation with extrinsic evaluation (2.4), we do not account for frequency effects (2.6) and polysemy (2.7). These aspects are left for future work.

6 Conclusion

This work explores an evaluation target for word embeddings: reaction time in lexical decision and naming tasks from a semantic priming experiment. Experiments show that this setting is dominated by different factors than word relatedness or similarity, and that the choice of algorithm is a stronger predictor of correlation than the choice of training corpora. In future work, we will leverage the non-word data from the lexical decision task in order to evaluate character-based word embeddings.

Acknowledgements Research supported by grants ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et ANR-11-IDEX-0001-02 (A*MIDEX).

[5]According to their section numbers

References

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

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *EMNLP*. Citeseer, pages 278–289.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* 00025.

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

Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*. pages 72–77.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276* .

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. ACM, pages 406–414.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *EMNLP*.

Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *of the Language Resources and Evaluation Conference (LREC 2016), Portoroz (Slovenia)*. pages 23–28.

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review* 23(6):1744–1756.

Keith A Hutchison, David A Balota, James H Neely, Michael J Cortese, Emily R Cohen-Shikora, Chi-Shing Tse, Melvin J Yap, Jesse J Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behavior Research Methods* 45(4):1099–1114.

Keith A Hutchison, James H Neely, and Jeffrey D Johnson. 2001. With great expectations, can two" wrongs" prime a" right"? *Journal of Experimental Psychology Learning Memory and Cognition* 27(6):1451–1463.

Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems* 31(6):5–14.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*. Citeseer, pages 302–308.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*. pages 171–180.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*. pages 104–113.

Timothy P McNamara. 2005. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

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

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

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology* 33(3-4):175–190.

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

Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. 2016. Exploring semantic representation in brain activity using word embeddings. In *EMNLP*. pages 669–679.

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

Tobias Schnabel, Igor Labutov, David M Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*. pages 298–307.

Anders Søgaard. 2016. Evaluating word embeddings with fmri and eye-tracking. *ACL 2016* page 116.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin* 87(2):245–251. 02852.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment .

Dongqiang Yang and David MW Powers. 2006. *Verb similarity on the taxonomy of WordNet*. Masaryk University.