Using word2vec to predict human language processing

Cassandra L. Jacobs (University of California, Davis; University of Toronto, Scarborough) &
Katrin Erk (University of Texas at Austin)
clxjacobs@ucdavis.edu

The greater the similarity between a word and its surrounding linguistic context, the faster readers move on to the next word. Typically, the similarity between words has been quantified by Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), a corpus-based measure of semantics. More recently, the **word2vec** algorithm has been shown to better predict semantic processing than LSA (e.g. Mandera, Keuleers, & Brysbaert, 2016), but little work has explored word2vec for sentence processing and there is little transparency as to model selection or design. The present work tests what model and data factors matter most for psycholinguists training their own word2vec models. We manipulated model parameters and the data source (text genre) to predict reading times (first fixation durations) from the Provo Corpus of Luke and Christianson (2016). In addition to the eyetracking data, each word in each sentence has an associated **cloze probability** from a word-by-word cloze task. Following Luke and Christianson (2016), we compare the semantic similarity of the **modal response** (the most common word predicted in the cloze task; e.g. *cat*) to the observed word (e.g. *dog*) for each model. We train our models on the Corpus of Contemporary American English (COCA, years 1990-2015; Davies, 2008) to more accurately approximate the linguistic knowledge of the participants in Luke and Christianson (2016).

word2vec. We compare and contrast several factors. First, we varied the two algorithms referred to as word2vec. The first, known as continuous bag-of-words (**CBOW**), predicts a missing word from its context. The other, **skip-gram**, predicts contexts from a single word. We train these models using the **gensim** Python implementation, which requires users to make a few important design decisions. We varied the number of surrounding words used to define a "context" (5 or 15 words) for training these models. We also tested whether different genres in COCA (Fiction, News, Spoken, Magazine, Academic, and a "Random 1 Million" sentences, and "all genres") varied in their ability to predict reading times.

Analyses. We analyzed (log) first fixation duration (FFD) as a function of the semantic similarity between the modal response (i.e. *cat*) at a given position and the word participants actually read at that position (i.e. *dog*) using linear mixed effects models with participant and word random intercepts. Models were compared using Chi-square tests over AIC. The greater the semantic similarity between the modal response and the observed word, the shorter FFDs are. Furthermore, all word2vec models better predict FFDs than LSA. **Corpus genre.** The best single genre was the Spoken subcorpus, though including all 5 genres improved model fit to the data (Table 1). **Algorithm (skip-gram and CBOW).** Skip-gram generally provided a better fit to the data than CBOW (Table 2). Combining both CBOW and skip-gram models better explains reading times than either model alone (Table 3). **Context window size**. Larger context windows were better at predicting FFDs than narrower ones (Table 2). Including both the narrow (5 words on each side) and wide context (15 words to each side) similarity scores best predicted FFDs above all models (Table 3).

Conclusions. word2vec shows considerable improvements over LSA in predicting human language processing performance. We have shown that different models all have slightly different predictive power for reading times, and that combined models typically perform best. The success of combined models suggests that researchers should combine multiple word2vec models for modeling semantic processing tasks unless they have clear reasons for selecting only one genre or model parameterization. More work must still be done to understand the locus of these semantic priming effects. However, the multifaceted nature of semantic processing means that researchers should aspire to capture this variability when using corpus measures for the design and analysis of their studies.

Table 1. Effect of genre on predicting reading times. The best model weights each genre separately (ensembled genres). *** = p < .001.

	AIC	R^2	Significance
LSA	93067	0.0037	-
Random 1 million	93027	0.0050	***
News	93016	0.0053	***
Magazine	93016	0.0053	-
Academic	93014	0.0054	***
Fiction	93014	0.0054	-
Spoken	92994	0.0060	***
All genres	92937	0.0077	***

Table 2. The effect of algorithm on predicting reading times trained on a random sample of 1 million sentences. The best single model uses skip-gram with a context of 15. *** = p < .001.

	AIC	R^2	Significance
Context = 5			
CBOW	93027	0.0050	***
Skip-gram	93037	0.0047	
Context = 10			
CBOW	93029	0.0049	
Skip-gram	93019	0.0052	***
Context = 15			
CBOW	93032	0.0048	
Skip-gram	93009	0.0055	***

Table 3. The effect of combining algorithms and models to predict reading times. Combining models and genres improves performance. * = p < .05, ** = p < .01, *** = p < .001.

	AIC	R^2	Significance
Genre			
Best genre (spoken)	92994	0.0060	
All genres	92937	0.0077	***
Algorithm			
Skip-gram	93009	0.0055	
Skip-gram + CBOW	93002	0.0059	**
Context window size			
15	93009	0.0055	
5 + 15	92968	0.0070	***

References

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159-190.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

Landauer & Dumais (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Luke, S. G. & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60.