USING WORD2VEC TO PREDICT HUMAN LANGUAGE PROCESSING

CASSANDRA L. JACOBS[1,2], KATRIN ERK[3]
1: UNIVERSITY OF TORONTO; 2: UNIVERSITY OF CALIFORNIA, DAVIS; 3: UNIVERSITY OF TEXAS AT AUSTIN

MOVING BEYOND LATENT SEMANTIC ANALYSIS

- LANGUAGE PROCESSING IS FASTER AND MORE ACCURATE WHEN A WORD'S MEANING MATCHES THE CONTEXT
- WORD2VEC (MIKOLOV ET AL., 2013) BEATS MANY OTHER COMPUTATIONAL MEASURES OF WORD MEANING
  - SKIP-GRAM
  - CONTINUOUS BAG-OF-WORDS (CBOW)

METHODS

- TRAIN MANY WORD2VEC MODELS ON DIFFERENT (SUB-) CORPORA AND MANIPULATING MODEL STRUCTURE
- GENRES FROM COCA CORPUS
  - E.G. FICTION, NEWS, MAGAZINES
- CONTEXT WINDOW SIZE (5, 10, 15)
- NEGATIVE SAMPLING (0, 5)
- ALGORITHM (SKIP-GRAM VS. CBOW)
- COMBINATIONS OF PREDICTORS FROM MANY WORD2VEC MODELS

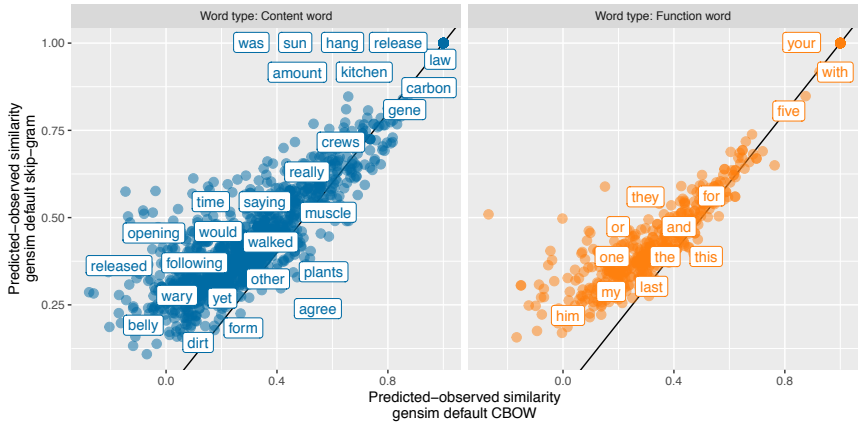

FIGURE 1: WORD SIMILARITY SCORES ARE HIGHLY SIMILAR ACROSS CBOW (X) AND SKIP-GRAM (Y) MODELS.

LUKE AND CHRISTIANSON (2016) EYE TRACKING DATA

- EYE TRACKING TASK WITH NATURALISTIC SENTENCES
- EVALUATION: PREDICTED-READ WORD SIMILARITY SCORES
- DEPENDENT MEASURES: FIRST FIXATION AND GAZE DURATION
  - DIFFERENTIALLY SENSITIVE TO SEMANTICS
- COVARIATES:
  - PARTICIPANT & ITEM RANDOM EFFECTS
  - WORD FREQUENCY
  - CLOZE UNCERTAINTY (ENTROPY)

RESULTS

- COMBINATIONS OF GENRES ✓✓✓
- COMBINATIONS OF CONTEXT WINDOW SIZES ✓✓✓
- NEGATIVE SAMPLING ✓✓✓
- SKIP-GRAM > CBOW FOR GAZE DURATION, OPPOSITE FOR FFD
- LARGER AND SMALLER WINDOW SIZE PERFORM SIMILARLY WELL

CONCLUSIONS

- GAZE DURATION AND FIRST FIXATION DURATION SENSITIVE TO SLIGHTLY DIFFERENT FACTORS
- COMBINED SIMILARITY MEASURES BEST EXPLAIN READING TIME DATA, OUTPERFORM LATENT SEMANTIC ANALYSIS
- PSYCHOLINGUISTICS SHOULD MOVE BEYOND LSA TO DESIGN, CONTROL FOR SEMANTIC PROCESSING

REFERENCES:
LUKE, S. G., & CHRISTIANSON, K. (2016). LIMITS ON LEXICAL PREDICTION DURING READING. *COGNITIVE PSYCHOLOGY*, 88, 22-60.
MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., & DEAN, J. (2013). DISTRIBUTED REPRESENTATIONS OF WORDS AND PHRASES AND THEIR COMPOSITIONALITY. IN *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS* (PP. 3111-3119).