Redefining Context Windows for Word Embedding Models: An Experimental Study

Pierre Lison
plison@nr.no
SAMBA - Statistical analysis, machine learning & image analysis
Norsk Regnesentral, Oslo, Norway

Andrey Kutuzov
andreku@ifi.uio.no
Language Technology Group
University of Oslo, Norway

Research question

- Distributional semantic models learn vector representations of words through the **contexts** they occur in.
- *Question: how does the choice of context window affect the type of embeddings that are learned?*
- We present here a systematic analysis of context windows based on **four** hyper-parameters:
  1. The *maximum size* of the context window
  2. The *weighting scheme* of context words according to their distance to the focus word
  3. The *relative position* of the context window (symmetric, left or right side)
  4. The treatment *of linguistic boundaries* such as end-of-sentence markers

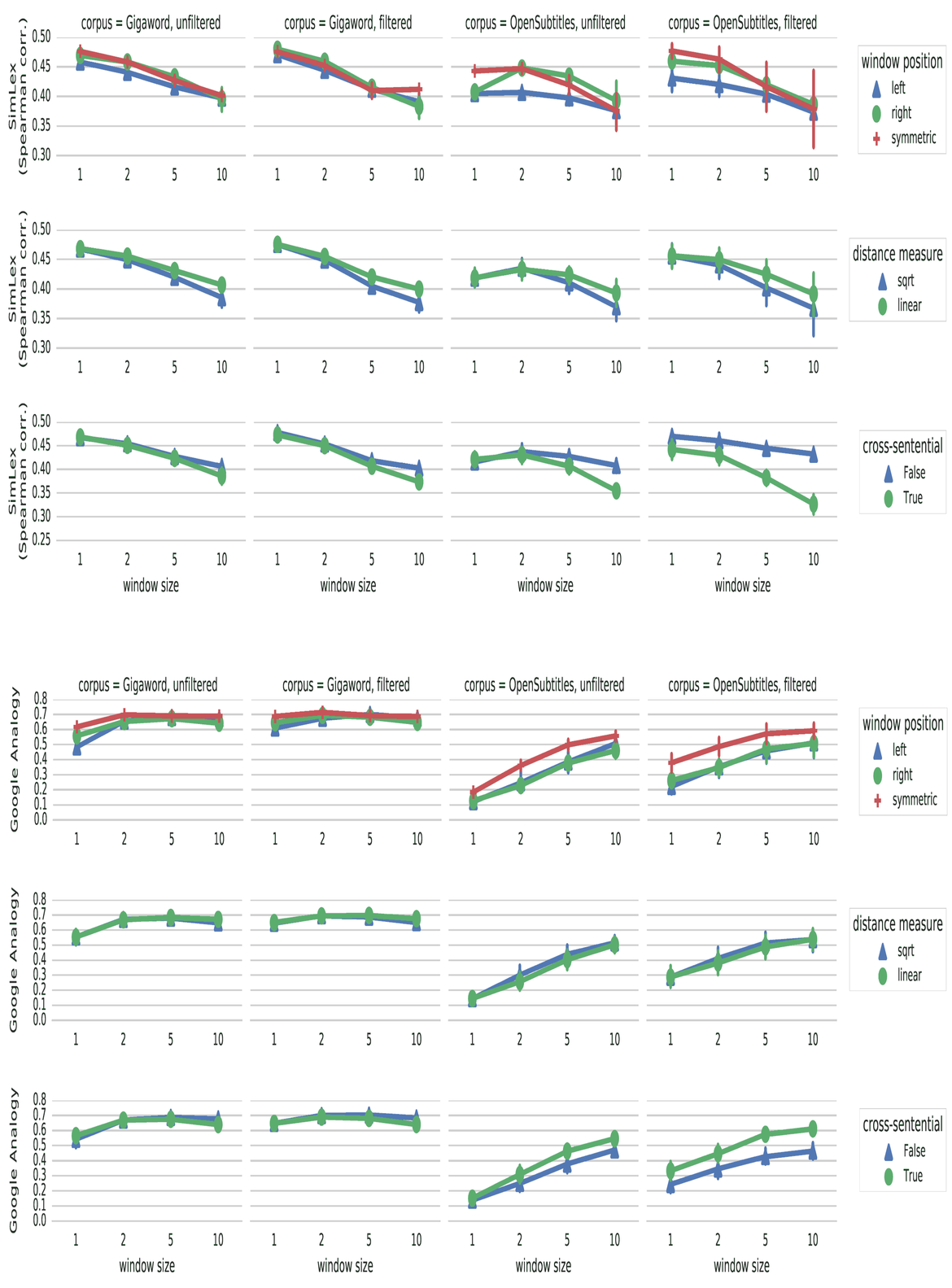
Background

- Distributional semantic models represent words through *real-valued vectors of fixed dimensions* based on the *distributional properties* of these words in large corpora.
- Latest generation of distributional models (*word2vec*, *GloVe*, etc.) can estimate *dense, low-dimensional vectors* called **embeddings** that capture various functional or topical relations between words.
- These models require the definition of a context for each word observed in a given corpus, often through a ***sliding window*** centered around the word to estimate.
- (But other types of contexts are possible, such as dependency-based or multilingual contexts)
- Context windows are defined by their **size**, their **weighting scheme** (e.g. the dynamic window mechanism in word2vec), their **position** and their **boundaries**

Experimental setup

- Embeddings trained using *Continuous Skip-gram with Negative Sampling* (SGNS) with 300-dimensional vectors, 10 negative samples per word and learned through 5 iterations.
- Two corpora used : *Gigaword* (4 billion tokens) and the English version of *OpenSubtitles* (700 million tokens), lemmatized & PoS-tagged. 2 versions:
- The results are computed with the Spearman correlation against the *SimLex-999* semantic similarity dataset and the accuracy on the semantic sections of the *Google Analogies Dataset*

Results



What we test

Hyperparameters tested:

1) Weighting scheme: **linear** or **squared**;
2) Max window size: **1, 2, 5, 10**
3) Window position: **left**, **right**, **symmetric**;
4) Cross-sentential boundaries: **True**, **False**;
5) Stop words removal before training: True (**filtered**), False (**unfiltered**).

All in all, 96 models trained and evaluated.

Main findings

1) **Cross-sentential contexts are beneficial for analogy tasks**, esp. for corpora with short sentences: sometimes a paragraph per line is better than a sentence per line!
2) For similarity tasks (at least in English), **right-side contexts** are much more important than the left-side contexts. The **window of *n* words to the right is almost as good as *n* words to the right AND to the left**.
3) *Word2vec* linear weighting scheme is a good choice.
4) Analogy task benefits from stop words removal.

NoDaLiDa 2017, Gothenburg, Sweden