Pretrained word and multi-sense embeddings for Estonian

Corpus: all embeddings are trained on lemmatized etTenTen: Corpus of the Estonian Web.

Word embeddings are trained with word2vec[1], [2], [3].

Sense embeddings are trained with SenseGram[4].

Sense inventory is induced from word embeddings.

Parameters

Models were trained using various parameter settings. The values of architecture, number of dimensions, window size, minimum frequency threshold and number of iterations vary, but other parameters follow default values declared here.

The exact values are declared in the name of the folders containing three output files:

.word_vectors - word vectors in the word2vec text format

.sense_vectors - sense vectors in the word2vec text format

.sense_vectors.inventory.csv - sense probabilities in TSV format

The **format** of the folder name is following: *architecture_dimensions_window_minc_iter.zip*:

- architecture: - CBOW or Skip-gram: *cbow* or *skip*;
- dimensions - number of dimensions: *100*, *150*, *300*, *450* or *750*;
- window - window size: *5*, *10*, *15* or *30*;
- minc - minimum count threshold: *2*, *5*, *10* or *15*;
- iter - number of iterations: *5*, *10* or *20*.

Download

architecture	dimensions	window size	mininum count	iterations	download
CBOW	100	5	10	20	cbow_100_5_10_20.zip
CBOW	150	15	10	20	cbow_150_15_10_20.zip
CBOW	150	15	5	20	cbow_150_15_5_20.zip
CBOW	150	5	10	20	cbow_150_5_10_20.zip
CBOW	150	5	10	5	cbow_150_5_10_5.zip
CBOW	150	5	5	20	cbow_150_5_5_20.zip
CBOW	300	10	10	5	cbow_300_10_10_5.zip
CBOW	300	15	10	20	cbow_300_15_10_20.zip
CBOW	300	15	10	5	cbow_300_15_10_5.zip
CBOW	300	1	10	20	cbow_300_1_10_20.zip

CBOW	300	30	10	20	cbow_300_30_10_20.zip
CBOW	300	5	10	10	cbow_300_5_10_10.zip
CBOW	300	5	10	20	cbow_300_5_10_20.zip
CBOW	300	5	10	5	cbow_300_5_10_5.zip
CBOW	300	5	15	5	cbow_300_5_15_5.zip
CBOW	300	5	2	20	cbow_300_5_2_20.zip
CBOW	300	5	5	20	cbow_300_5_5_20.zip
CBOW	300	5	5	5	cbow_300_5_5_5.zip
CBOW	450	5	10	5	cbow_450_5_10_5.zip
CBOW	750	5	10	20	cbow_750_5_10_20.zip
Skip-gram	150	5	10	5	skip_150_5_10_5.zip
Skip-gram	300	10	10	5	skip_300_10_10_5.zip
Skip-gram	300	15	10	5	skip_300_15_10_5.zip
Skip-gram	300	5	10	10	skip_300_5_10_10.zip
Skip-gram	300	5	10	20	skip_300_5_10_20.zip
Skip-gram	300	5	10	5	skip_300_5_10_5.zip
Skip-gram	300	5	15	5	skip_300_5_15_5.zip
Skip-gram	300	5	5	5	skip_300_5_5_5.zip
Skip-gram	450	5	10	5	skip_450_5_10_5.zip

Credits

Author: Eleri Aedmaa (Institute of Estonian and General Linguistics, University of Tartu)

This work was carried out in the High Performance Computing Center of University of Tartu.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

[3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.

[4] Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. Making Sense of Word Embeddings. In Proceedings of the 1st Workshop on Representation Learning for NLP, 2016.