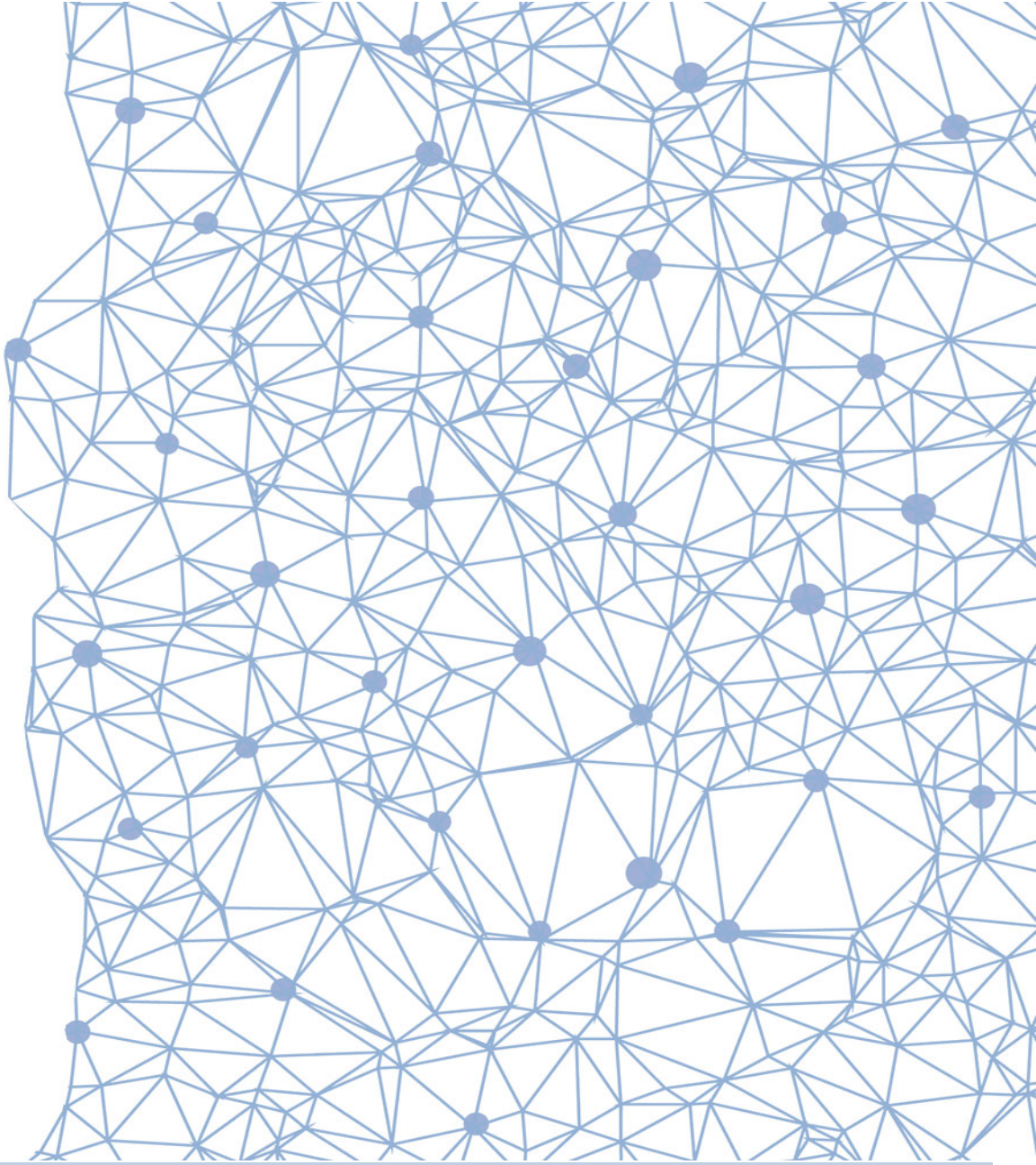
BITVILLE

NLP with Deep Learning

Antti Keurulainen Lic.Sc (Tech.)

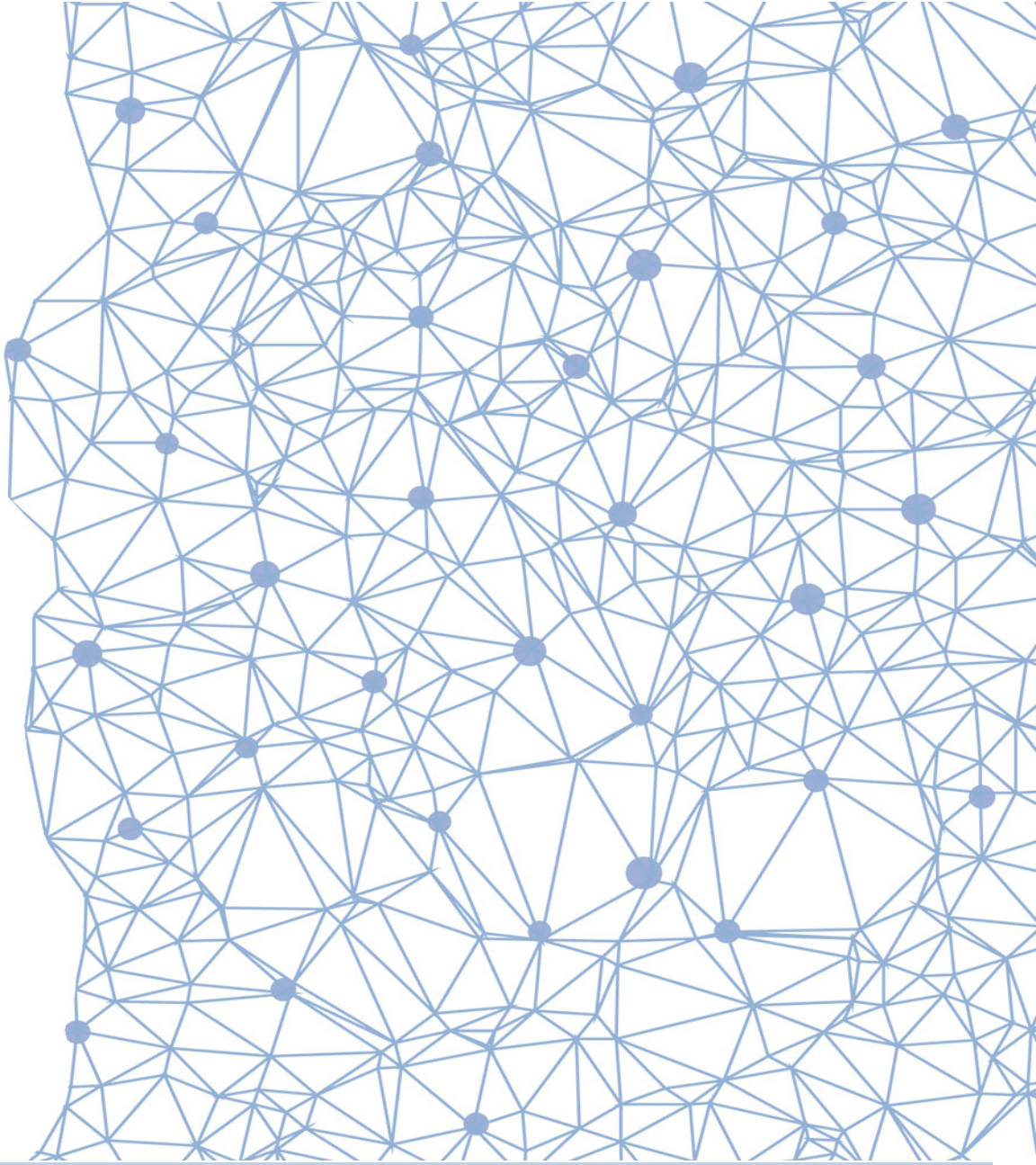
Member in Aalto University Probabilistic Machine Learning research group
Research director at Bitville AI

antti.keurulainen@bitville.com
@AnttiKeurulaine



Agenda

- Introduction to NLP
- Bag of Words
- Distributed word representations – word2vec
- Pretrained language models
- BERT
- GPT-2
- MT-DNN



Introduction to natural language processing (NLP)

Natural Language Processing

Wikipedia 2019:

”Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.”

Typical NLP tasks

- Sentiment analysis
- Text to speech
- Speech to text (speech recognition)
- Question answering
- Text summarization
- Machine transalation
- Named entity recognition
- Automated essay scoring
- …

Bag of words

[0 0 0 0 0 0 … 0 0 1 0 0 … 0 0 0 0]

Each word in the vocabulary has a one-hot representation

“This is the best movie I have ever seen”

“Peter and Mary like movies” [1 0 1 1 1 1 0 0]

“I like to watch movies” [0 1 1 0 1 0 1 1]

[1 1 2 1 2 1 1 1]

Vocabulary:
and
I
like
mary
movies
peter
to
watch

How to measure the similarity of the words?

No mechanism to describe the relationships between the words.

Information about the word order is lost

Bag of words

[0 0 0 0 0 0 … 0 0 1 0 0 … 0 0 0 0]

Each word in the vocabulary has a one-hot representation

“This is the best movie I have ever seen”

“Peter and Mary like movies” [1 0 1 1 1 1 0 0]

“I like to watch movies” [0 1 1 0 1 0 1 1]

[1 1 2 1 2 1 1 1]

Vocabulary:
and
I
like
mary
movies
peter
to
watch

How to measure the similarity of the words?

No mechanism to describe the relationships between the words.

Information about the word order is lost

Word embedding - distributed representation

You shall know a word by the company it keeps (Firth, J. R. 1957:11)

John Rupert Firth (June 17, 1890 in Keighley, Yorkshire – December 14, 1960 in Lindfield, West Sussex), commonly known as **J. R. Firth**, was an English linguist and a leading figure in British linguistics during the 1950s.[1] He was Professor of English at the University of the Punjab from 1919–1928. He then worked in the phonetics department of University College London before moving to the School of Oriental and African Studies, where he became Professor of General Linguistics, a position he held until his retirement in 1956.[2]

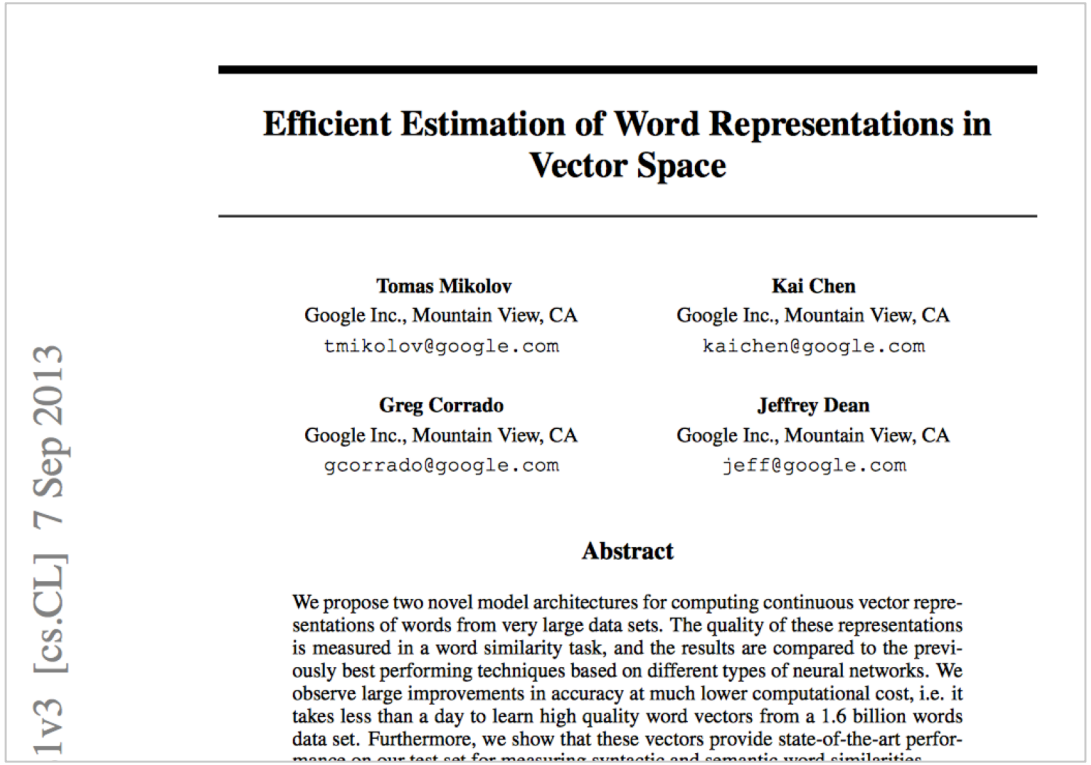
John R. Firth

Mikolov et al (Google 2013)

Efficient Estimation of Word Representations in Vector Space

The natural language words must be converted to a format that can be manipulated with a computer (i.e set of numbers).

This method makes the representation of a word based on what are the other words typically close to the word in question.



1v3 [cs.CL] 7 Sep 2013

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art perfor-

https://arxiv.org/pdf/1301.3781.pdf

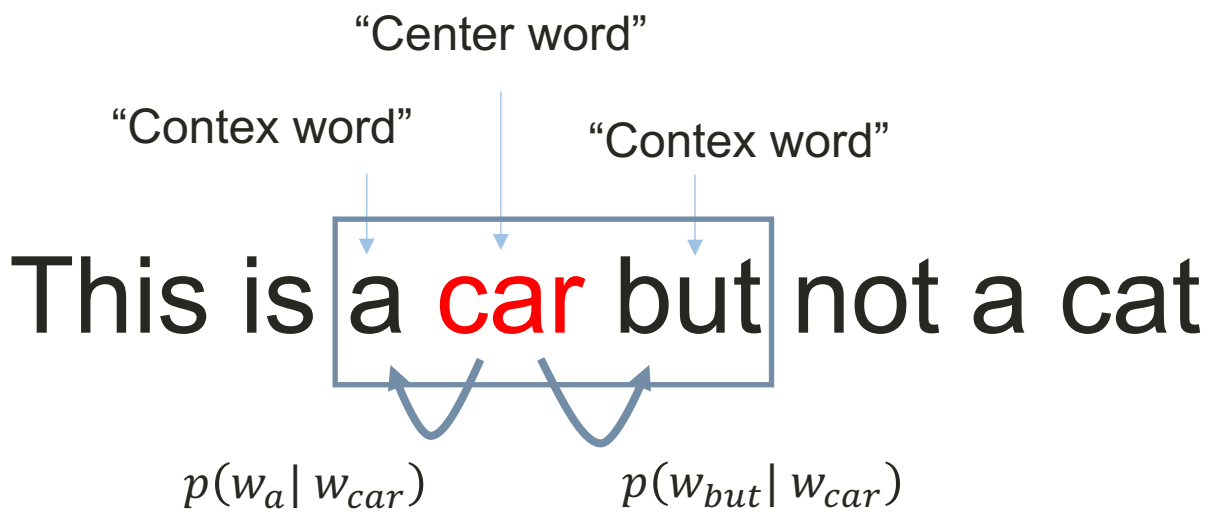
Word embeddings
(words in vector format)

Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

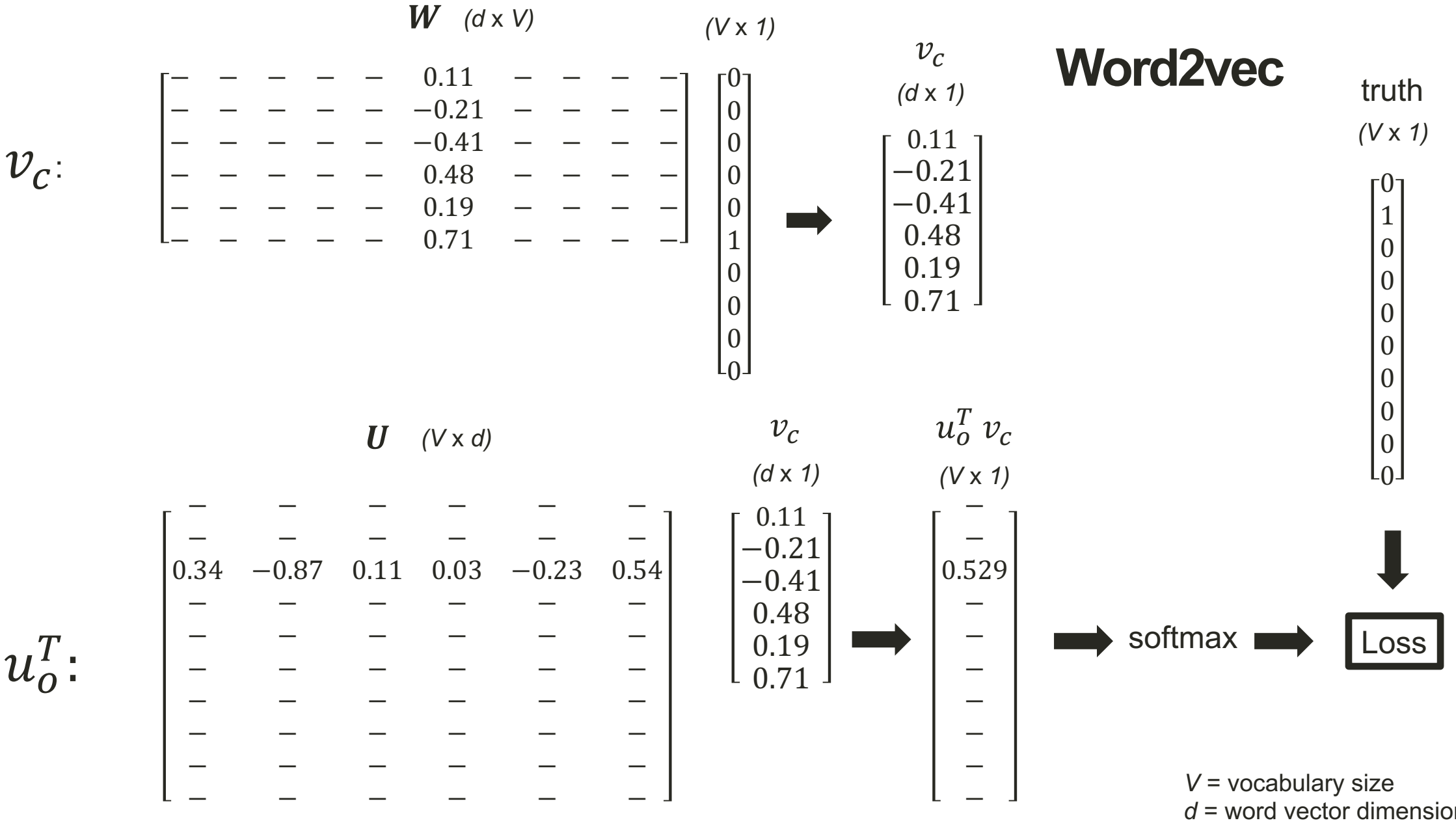
A step towards machine reasoning

Mikolov et al 2013:
https://www.aclweb.org/anthology/N/N13/N13-1090.pdf

Word embedding; distributed representation



Word2vec


v_c:
$\boldsymbol{W}$ (d x V)
(V x 1)
v_c
(d x 1)
truth
(V x 1)
u_o^T:
$\boldsymbol{U}$ (V x d)
v_c
(d x 1)
$u_o^T v_c$
(V x 1)
softmax
Loss
V = vocabulary size
d = word vector dimensionality

Pretrained language models

BERT

GPT-2

MT-DNN

Live demo

Pretrained language model