Improving the Compositionality of Word Embeddings

MASTER THESIS

Author:
Thijs SCHEEPERS

Supervisors:
dr. Evangelos KANOULAS
dr. Efstratios GAVVES

Truely understanding

A far out goal for Artificial Intelligence

What is your name?

Such a simple question

from *Her*
by Spike Jonze (2013)

“What is your name?”

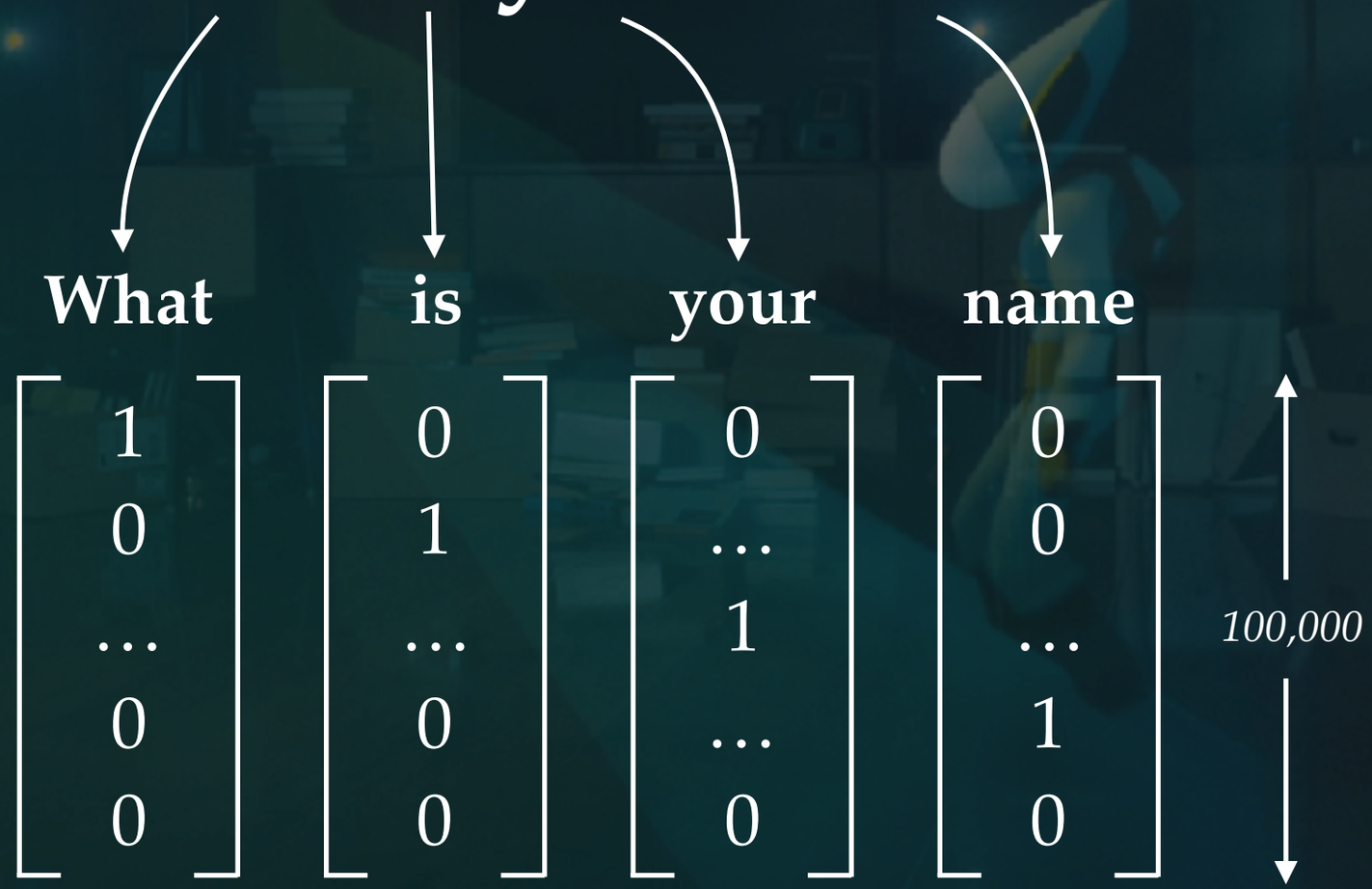
↓

01010111 01101000 01100001 01110100 00100000 01101001
01110011 00100000 01111001 01101111 01110101 01110010
00100000 01101110 01100001 01101101 01100101 00111111

Transforming to Binary


“What is your name?”
...
01010111 01101000 01100001 01110100 00100000 01101001
01110011 00100000 01111001 01101111 01110101 01110010
00100000 01101110 01100001 01101101 01100101 00111111
ASCII

“What is your name?”



Bag-of-words

Improving the Compositionality of **Word Embeddings**

Title of the Master Thesis

“What is your name?”

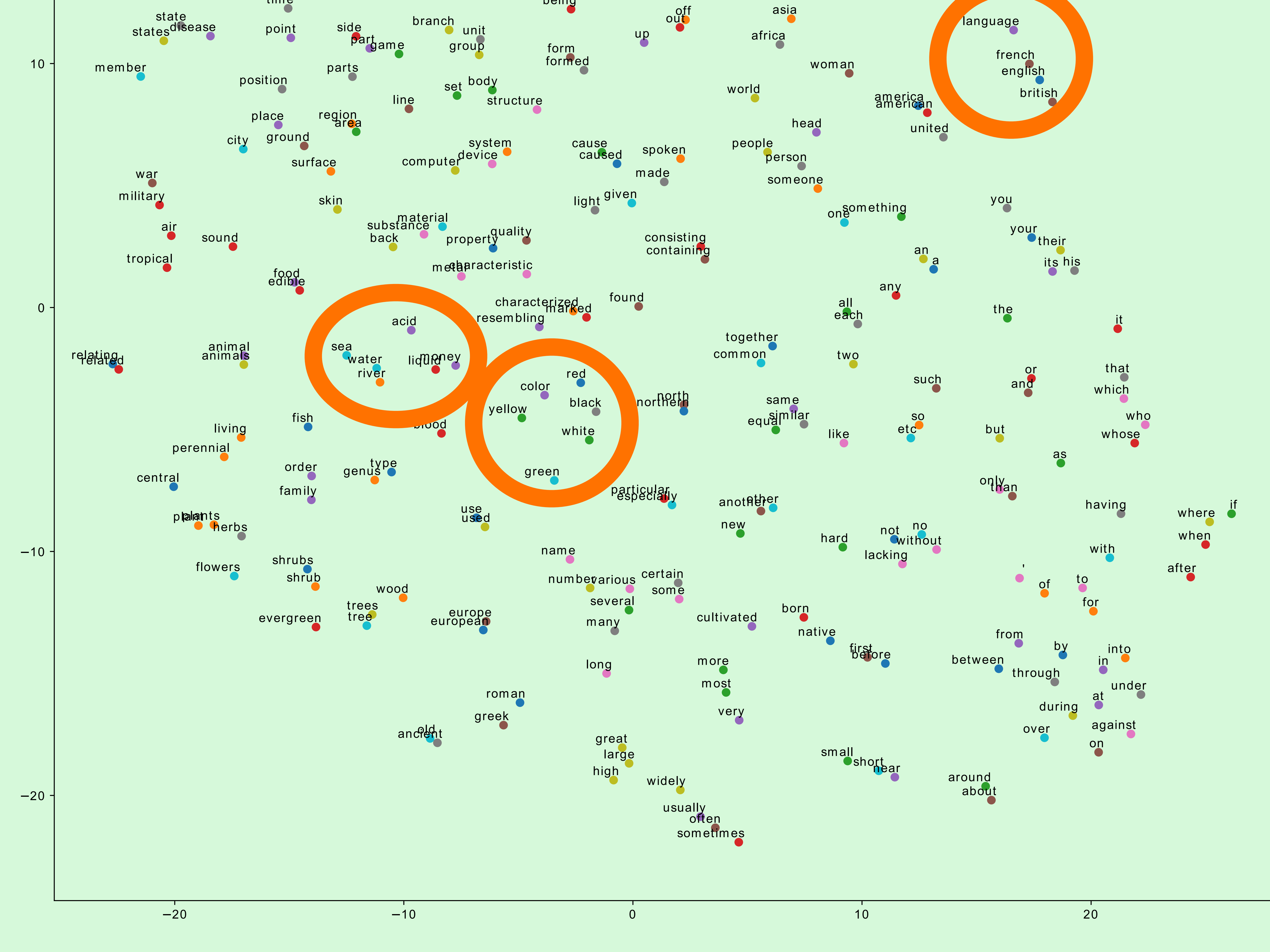
What	is	your	name	
0.23	1.62	-1.60	0.87	
1.56	-0.25	0.82	1.32	
…	…	…	…	*300*
-0.78	-0.53	0.91	-1.41	
0.93	1.72	-1.39	-0.91	

Word Embeddings

Word Embeddings encode Lexical Semantics, i.e. word meaning

What	**is**	**your**	**name**
0.23	1.62	-1.60	0.87
1.56	-0.25	0.82	1.32
…	…	…	…
-0.78	-0.53	0.91	-1.41
0.93	1.72	-1.39	-0.91

300

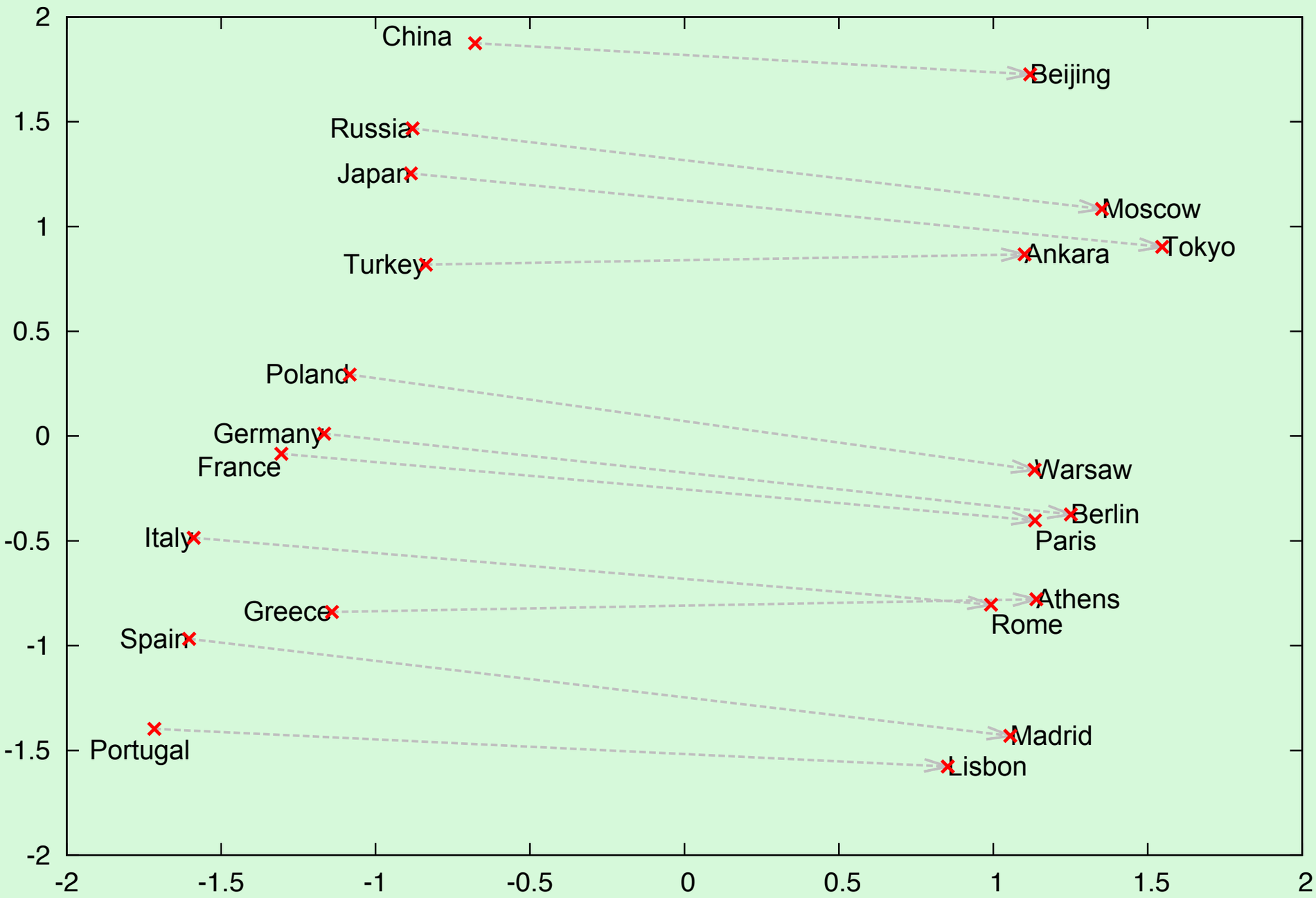


state
states
disease
point
side
part
game
branch
unit
group
being
off
out
up
asia
africa
language
french
english
british
member
position
parts
form
formed
woman
set
body
line
structure
world
america
american
place
region
area
city
ground
head
united
system
device
cause
caused
spoken
people
person
surface
computer
made
war
someone
military
skin
light
given
one
something
you
air
material
substance
sound
back
property
quality
consisting
containing
your
their
tropical
an
a
its
his
food
edible
metal
characteristic
any
all
each
found
characterized
marked
resembling
the
it
acid
together
sea
water
liquid
money
common
relating
related
animal
animals
two
river
red
color
that
or
such
and
which
north
northern
black
yellow
same
similar
who
fish
equal
so
blood
living
white
like
etc
but
whose
perennial
as
order
type
genus
green
central
family
particular
especially
only
than
another
other
having
use
used
where
if
plants
plant
herbs
new
no
not
without
hard
when
name
lacking
with
shrubs
flowers
shrub
certain
number
various
some
of
to
after
wood
several
for
trees
tree
europe
european
born
evergreen
many
cultivated
native
from
by
into
first
before
between
in
more
through
long
under
most
roman
at
during
very
greek
against
old
ancient
over
great
on
large
small
short
near
high
widely
around
about
usually
often
sometimes
10
0
−10
−20
20

Word Embedding space

$$\tfrac{1}{5} \cdot (\text{‘Berlin’} - \text{‘Germany’}) +$$
$$(\text{‘Stockholm’} - \text{‘Sweden’}) +$$
$$(\text{‘Washington DC’} - \text{‘United States’}) +$$
$$(\text{‘Beijing’} - \text{‘China’}) +$$
$$(\text{‘London’} - \text{‘United Kingdom’}) \approx \{\text{capital}\}$$

$$\text{‘Netherlands’} + \{\text{capital}\} = \text{‘Amsterdam’}$$



from Mikolov et al. (2013)

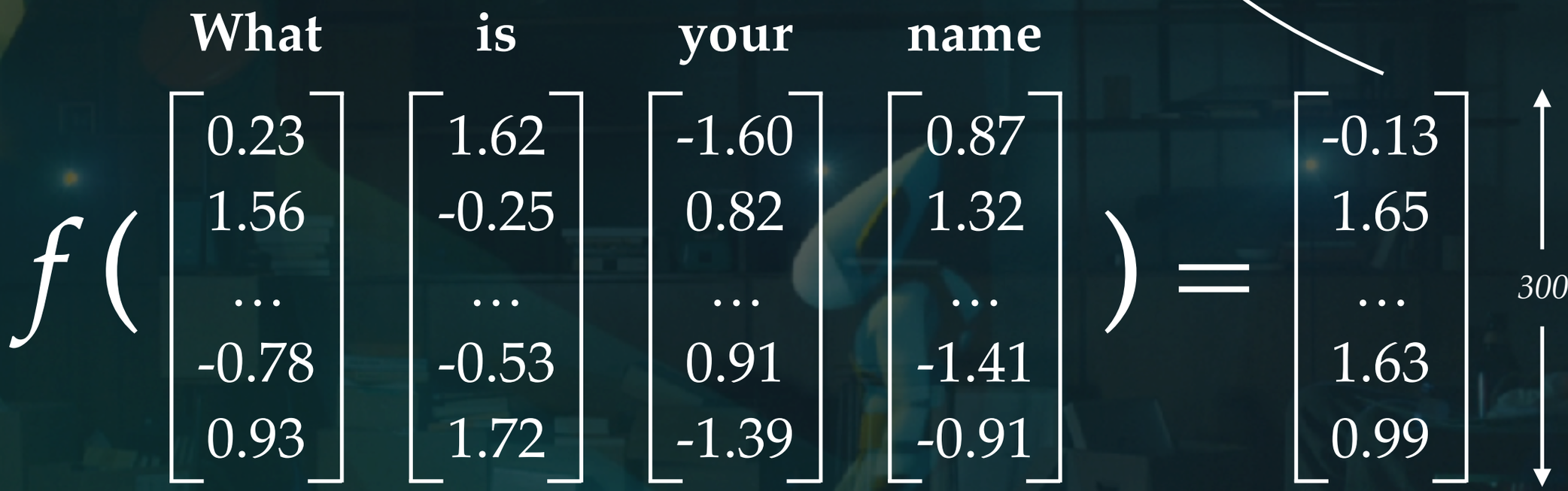
Improving the **Compositionality** of Word Embeddings

Title of the Master Thesis

Word Embedding Composition

Combine encodings of *word meanings* in such a way that a good encoding of *their joint meaning* is created

“What is your name?”



Word Embedding Composition

Overview

1. Evaluating *compositionality*
2. Tuning word embeddings for better *algebraic composition*
3. *Neural methods* for composing word embeddings

1. Evaluating compositionality

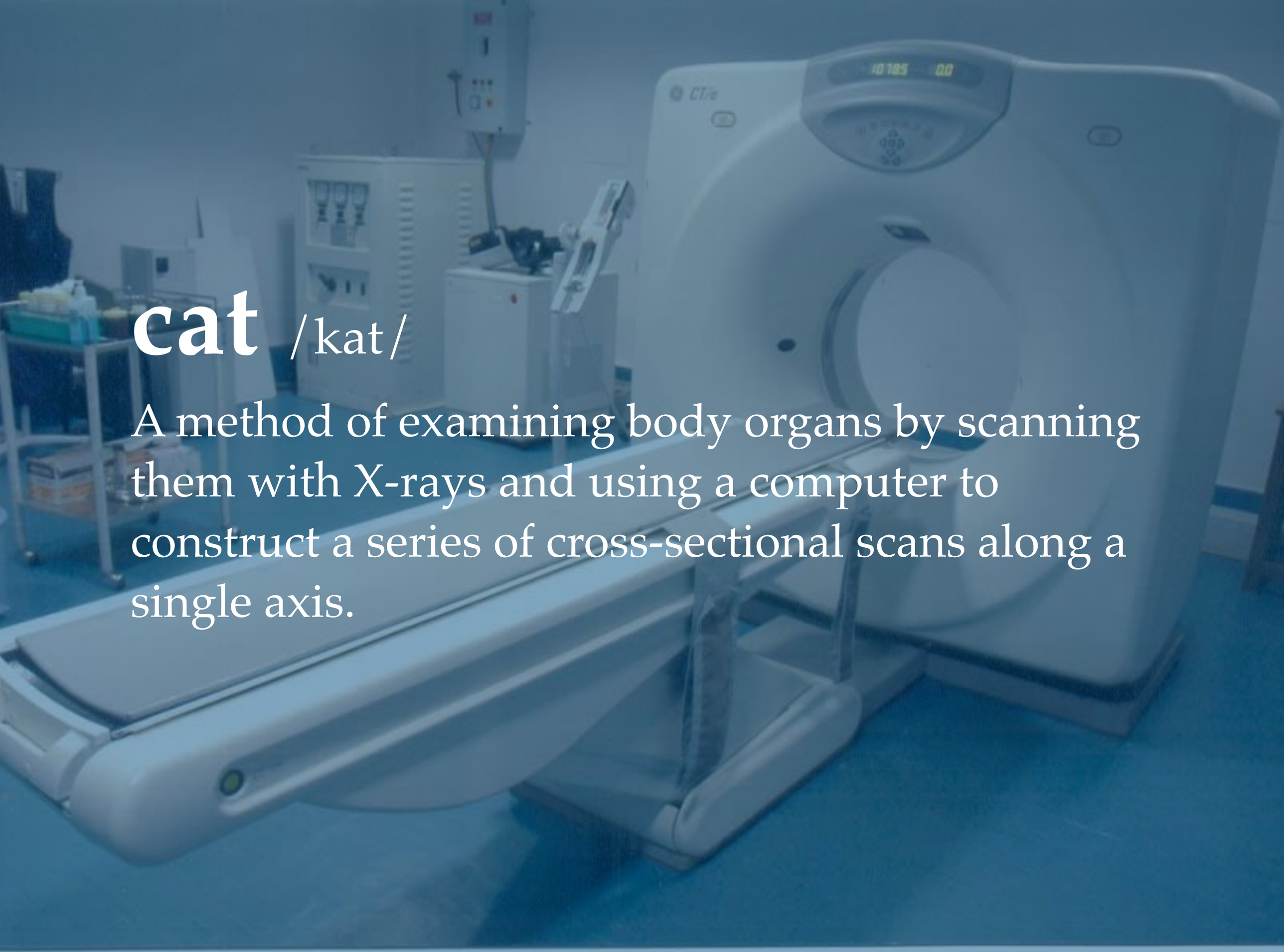
Introducing *CompVecEval* a method to evaluate word embeddings on their compositionality

Dictionaries

A pragmatic solution for *word meaning*

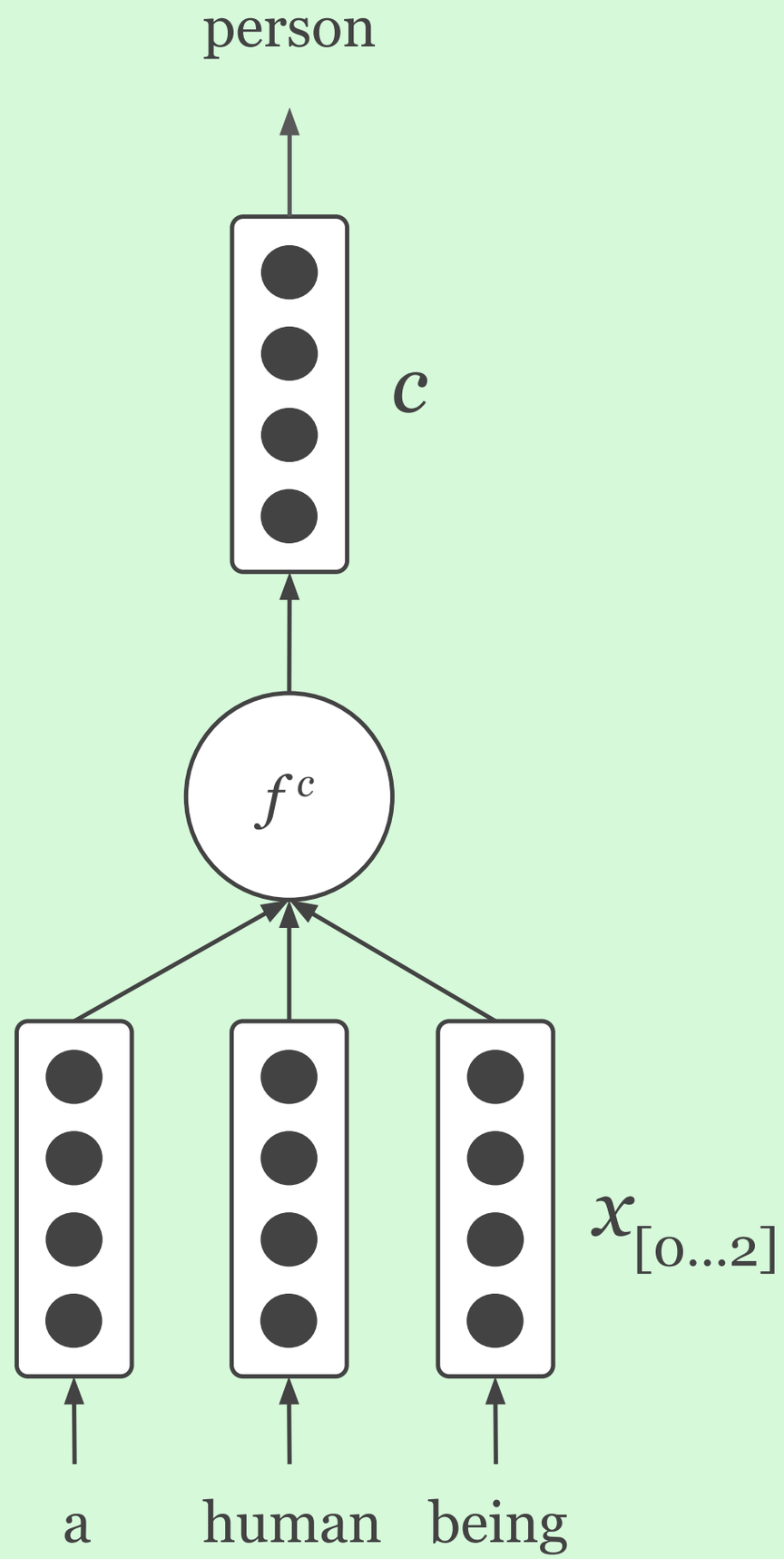
cat /kat/

A small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.



cat /kat/

A method of examining body organs by scanning them with X-rays and using a computer to construct a series of cross-sectional scans along a single axis.


person
c
f^c
$x_{[0...2]}$
a
human
being

Dictionary

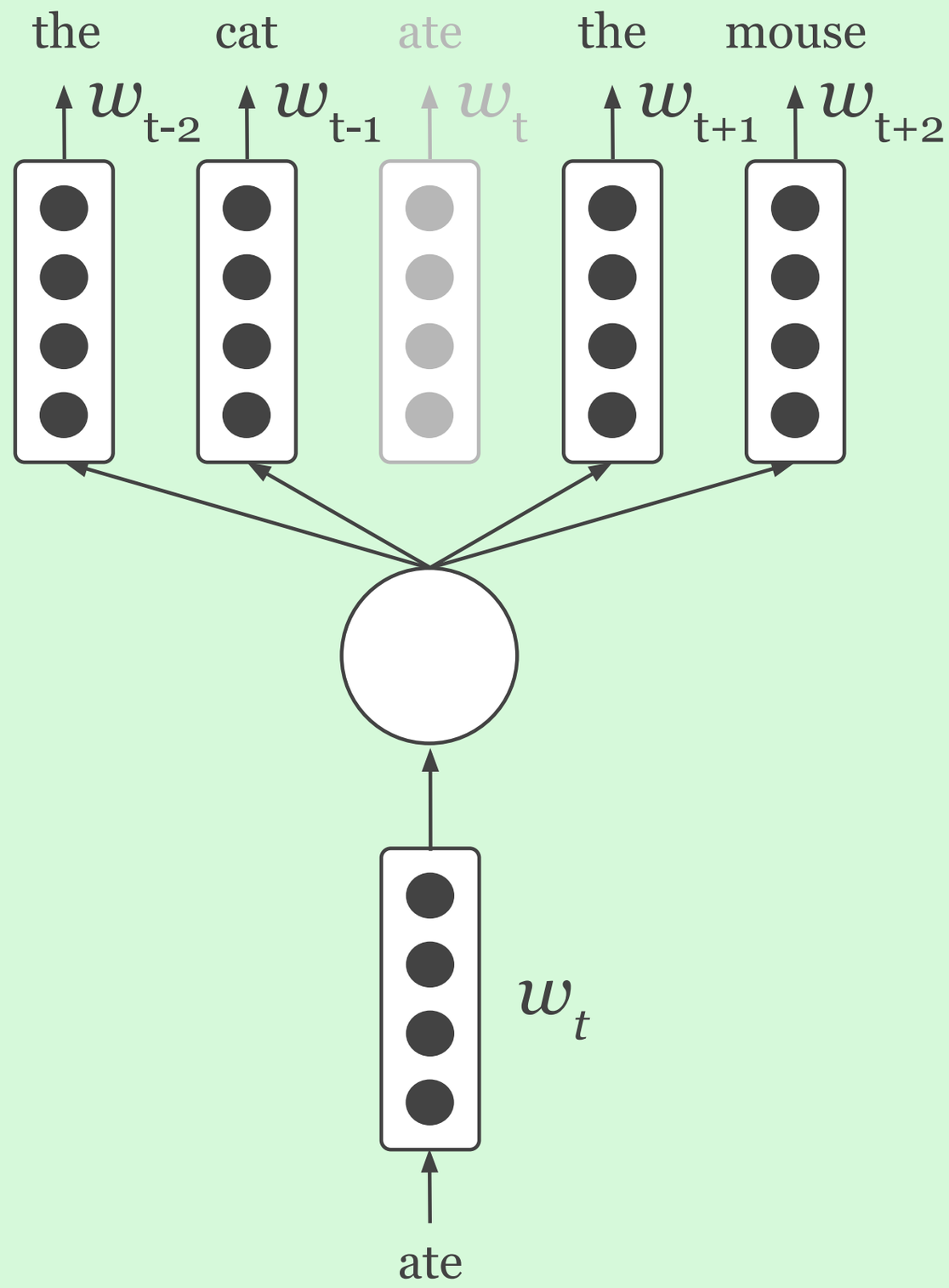
1. WordNet (Miller and Fellbaum 1998)

2. We use *4,119* datapoints for our evaluation method, and *72,322* datapoints for tuning

Popular pretrained Word Embeddings

1. Word2Vec (Mikolov et al. 2013)
2. GloVe (Pennington et al. 2014)
3. fastText (Bojanowski et al. 2016)
4. Paragram (Wieting et al. 2015)

Word2Vec Skip-gram



Additive Compositionality proven for Skip-Gram*

1. Uniform distribution is assumed
2. Definition of compositionality

Skip-Gram – Zipf + Uniform = Vector Additivity

Alex Gittens
Dept. of Computer Science
Rensselaer Polytechnic Institute
gittea@rpi.edu

Dimitris Achlioptas
Dept. of Computer Science
UC Santa Cruz
optas@soe.ucsc.edu

Michael W. Mahoney
ICSI and Dept. of Statistics
UC Berkeley
mmahoney@stat.berkeley.edu

Abstract

In recent years word-embedding models have gained great popularity due to their remarkable performance on several tasks, including word analogy questions and caption generation. An unexpected "side-effect" of such models is that their vectors often exhibit compositionality, i.e., *adding* two word-vectors results in a vector that is only a small angle away from the vector of a word representing the semantic composite of the original words, e.g., "man" + "royal" = "king".

This work provides a theoretical justification for the presence of additive compositionality in word vectors learned using the Skip-Gram model. In particular, it shows that additive compositionality holds in an even stricter sense (small distance rather than small angle) under certain assumptions on the process generating the corpus. As a corollary, it explains the success of vector calculus in solving word analogies. When these assumptions do not hold, this work describes the correct non-linear composition operator.

Finally, this work establishes a connection between the Skip-Gram model and the Sufficient Dimensionality Reduction (SDR) framework of Globerson and Tishby: the parameters of SDR models can be obtained from those of Skip-Gram models simply by adding information on symbol frequencies. This shows that Skip-Gram embeddings are optimal in the sense of Globerson and Tishby and, further, implies that the heuristics commonly used to approximately fit Skip-Gram models can be used to fit SDR models.

1 Introduction

The strategy of representing words as vectors has a long history in computational linguistics and machine learning. The general idea is to find a map from words to vectors such that word-similarity and vector-similarity are in correspondence. Whilst vector-similarity can be readily quantified in terms of distances and angles, quantifying word-similarity is a more ambiguous task. A key insight in that regard is to posit that the meaning of a word is captured by "the company it keeps" (Firth, 1957) and, therefore, that two words that keep company with similar words are likely to be similar themselves.

In the simplest case, one seeks vectors whose inner products approximate the co-occurrence frequencies. In more sophisticated methods co-occurrences are reweighed to suppress the effect of more frequent words (Rohde et al., 2006) and/or to emphasize pairs of words whose co-occurrence frequency maximally deviates from the independence assumption (Church and Hanks, 1990).

An alternative to seeking word-embeddings that reflect co-occurrence statistics is to extract the vectorial representation of words from non-linear statistical language models, specifically neural networks. (Bengio et al., 2003) already proposed (i) associating with each vocabulary word a *feature vector*, (ii) expressing the *probability function* of word sequences in terms of the feature vectors of the words in the sequence, and (iii) learning *simultaneously* the vectors and the parameters of the probability function. This approach came into prominence recently through works of Mikolov et al. (see below) whose main departure from (Bengio et al., 2003) was to follow the suggestion of (Mnih and Hinton, 2007) and trade-away the expressive capacity of general neural-network models for the scalability (to very large

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 69–76
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1007

Evaluating by Ranking

1. We rank lemmas according to their euclidean distance
2. We use a ball-tree algorithm to make this efficient
3. We considered several ranking metrics, and choose to use Mean Reciprocal Rank

Algebraic Composition

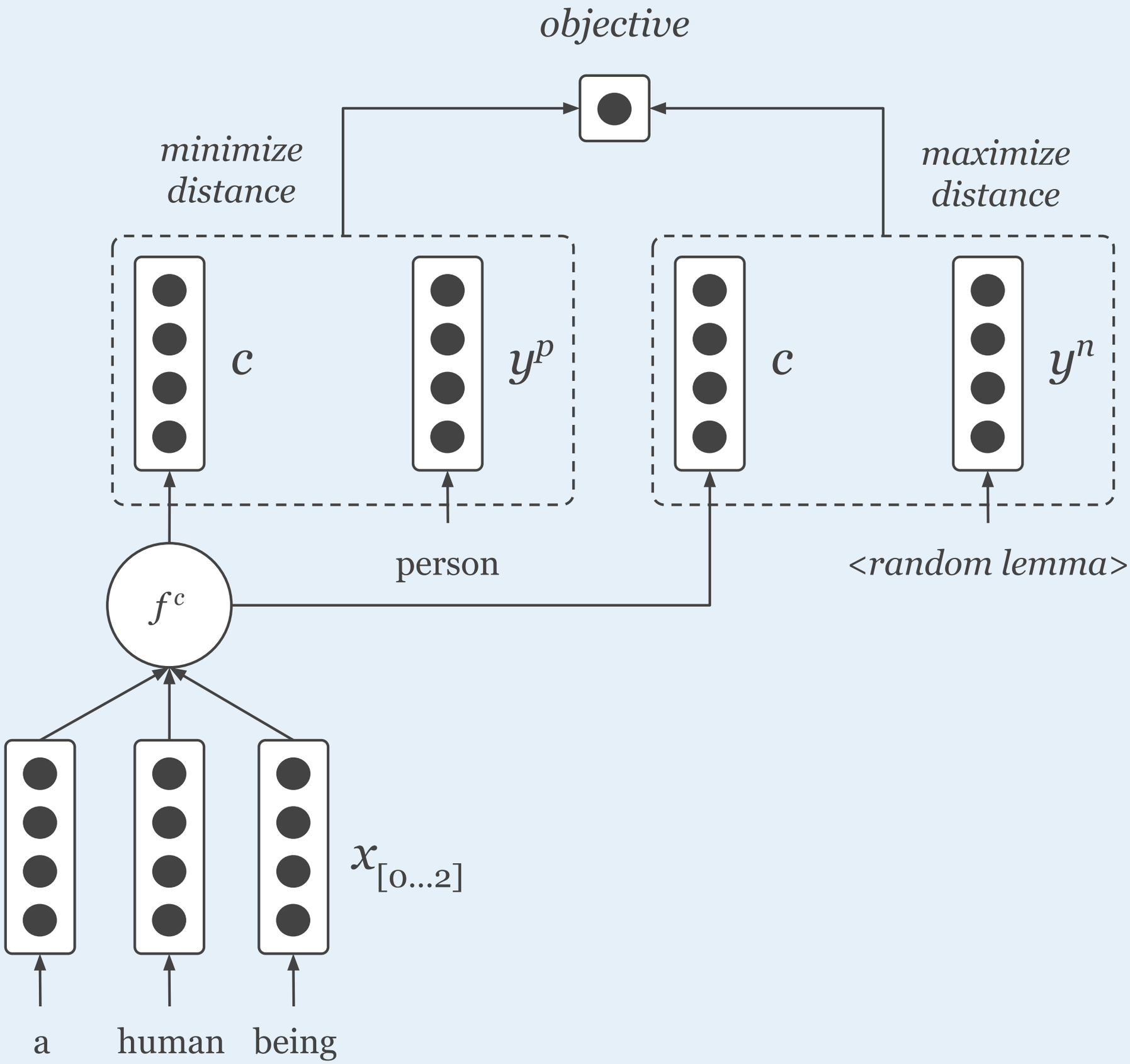
1. Addition
2. Averaging
3. Multiplication
4. Max-pooling

Evaluation Results

	Word2Vec	GloVe	fastText	Paragram
random	0.7 %			
+	16.8 %	11.9 %	20.7 %	**26.5 %**
avg(d)	2.0 %	3.3 %	3.0 %	*3.8 %*
×	0.6 %	0.9 %	0.9 %	*1.0 %*
max(d)	6.6 %	13.7 %	14.6 %	*20.5 %*

2. Tuning for algebraic composition

Improving existing *word embeddings* by tuning them to algebraically compose *lexicographic data*

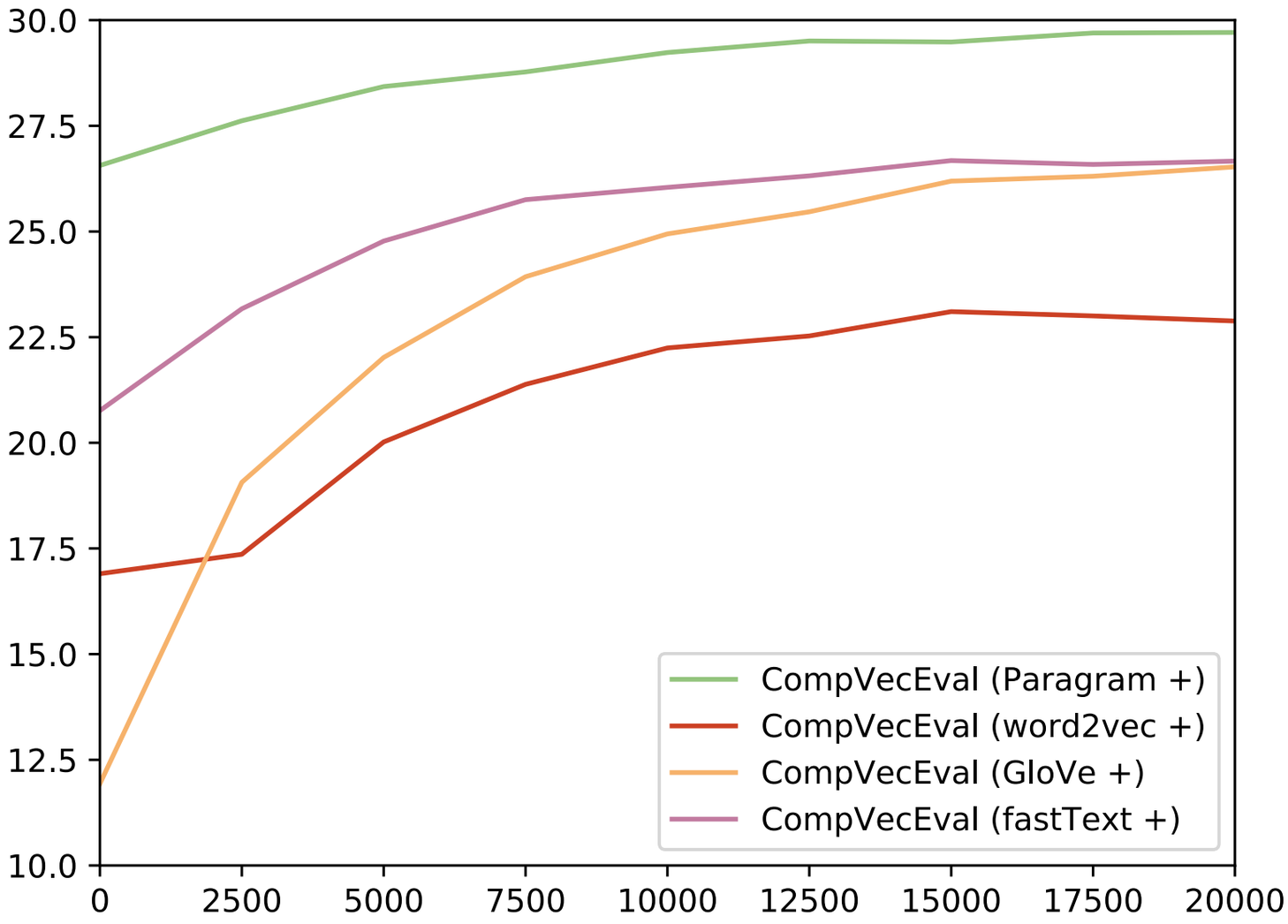

objective
minimize distance
maximize distance
c
y^p
c
y^n
person
<random lemma>
f^c
$x_{[0...2]}$
a
human
being

Objective Function

1. Triplet loss function
2. Negative example
3. Within a margin

$$\text{triplet loss} := \sum_{i=1}^{N} \max\Big(||c_i - y_i^p||^2 - ||c_i - y_i^n||^2 + \alpha, 0 \Big)$$

CompVecEval

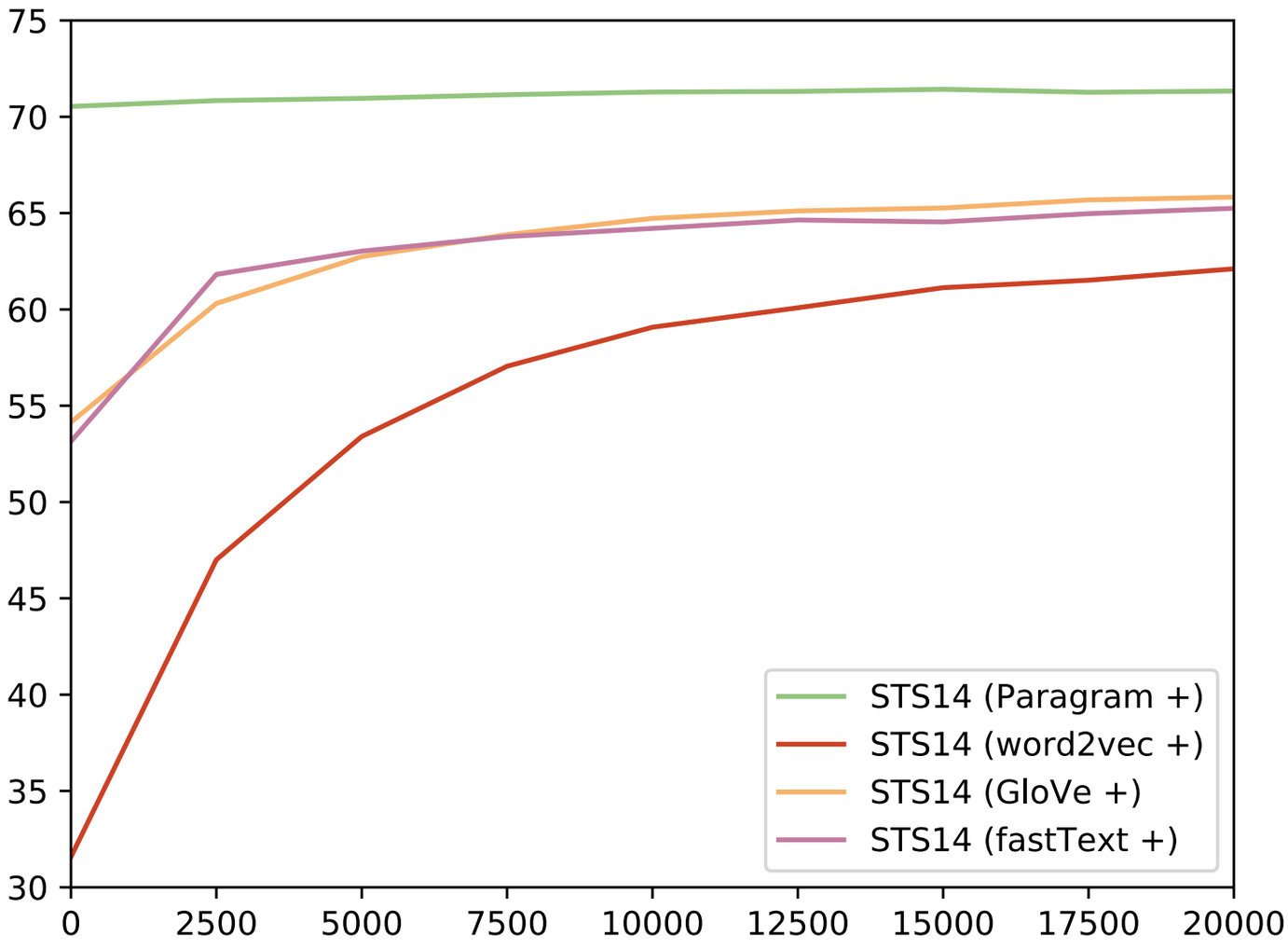


Evaluating tuned embeddings

1. We evaluated using *CompVecEval*
2. But also using 15 existing *sentence representation evaluation* methods
3. And 13 existing *word representation evaluation* methods

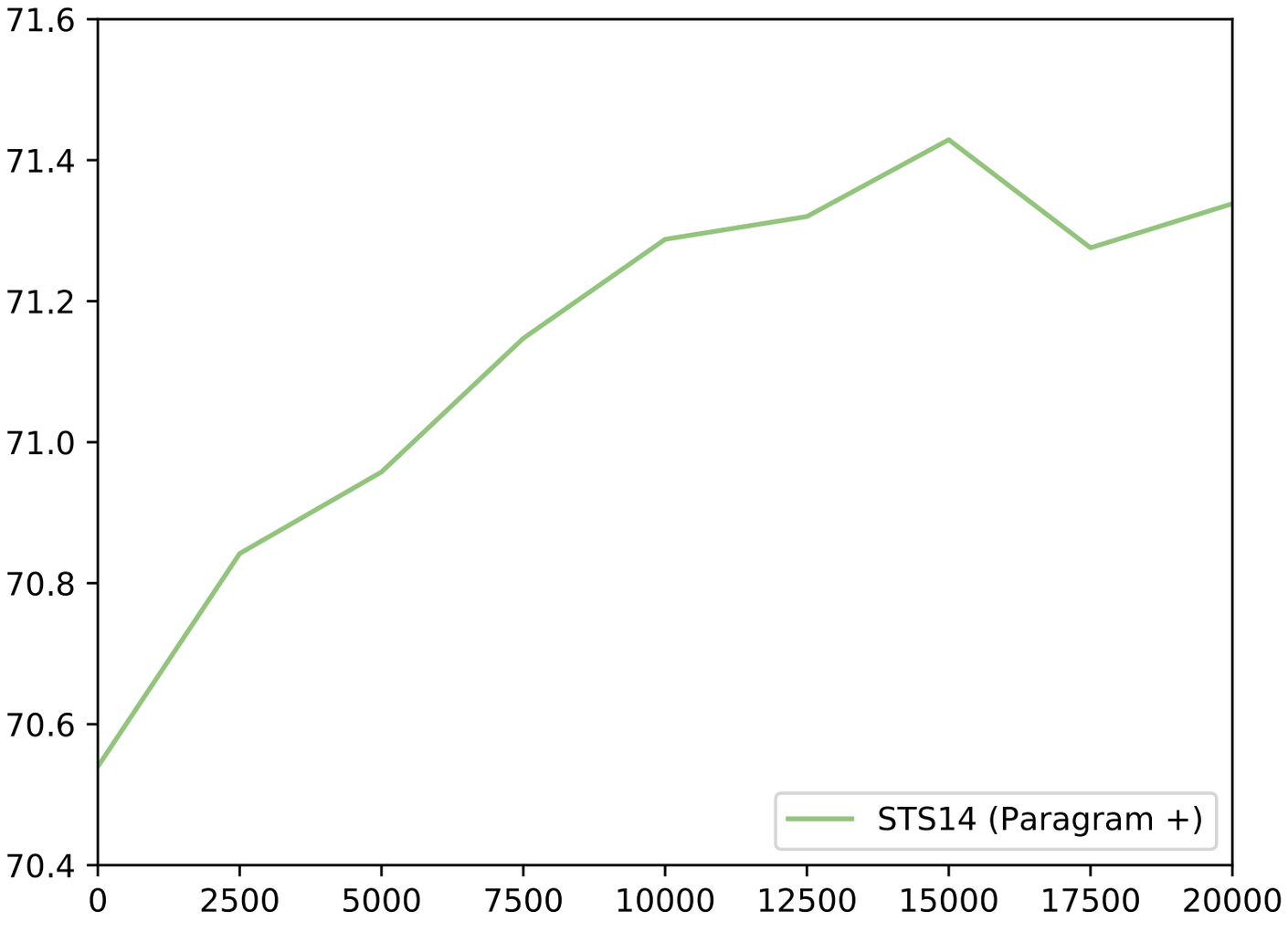
STS14

Sentence representation evaluation



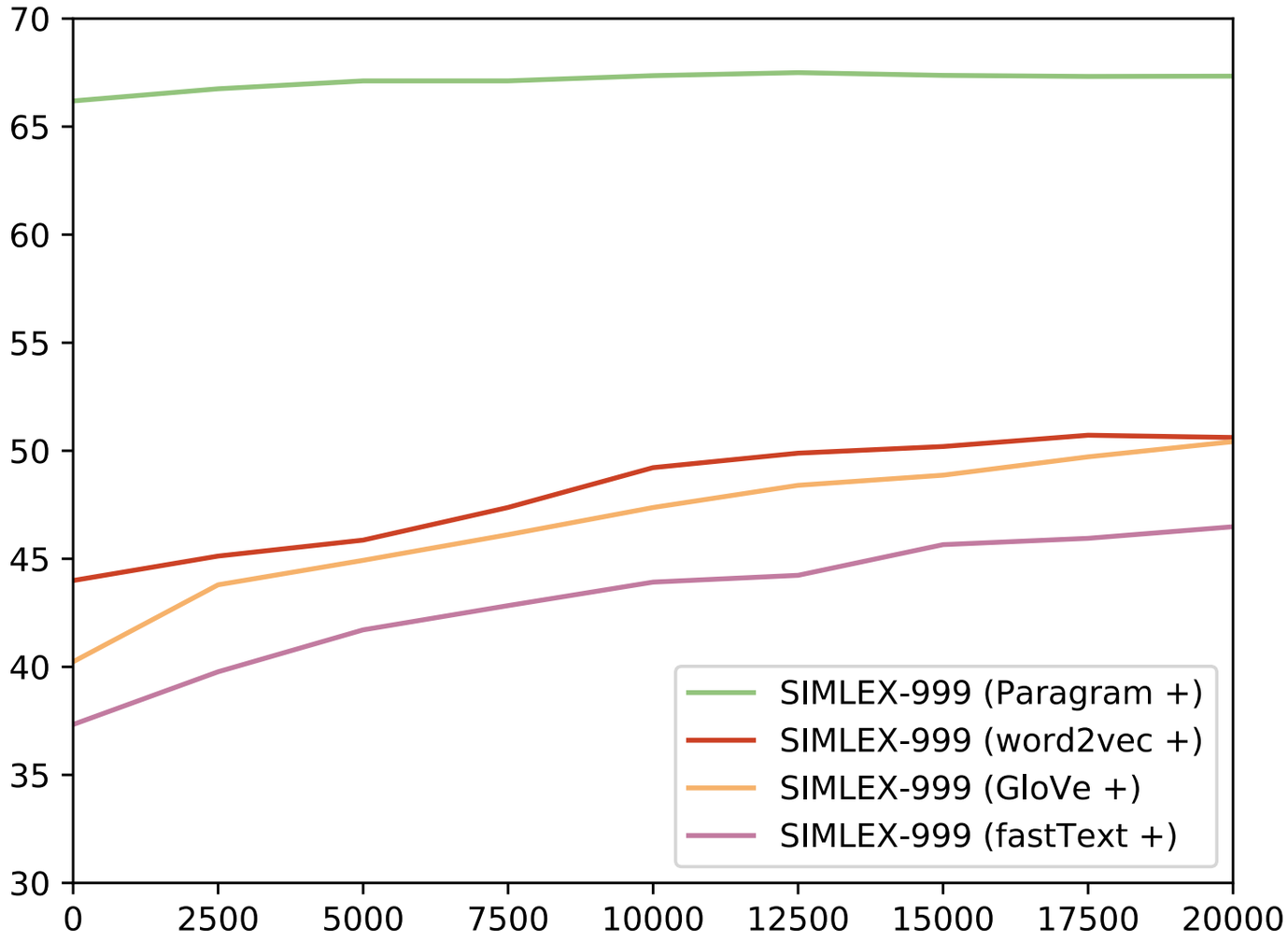
STS14

Sentence representation evaluation



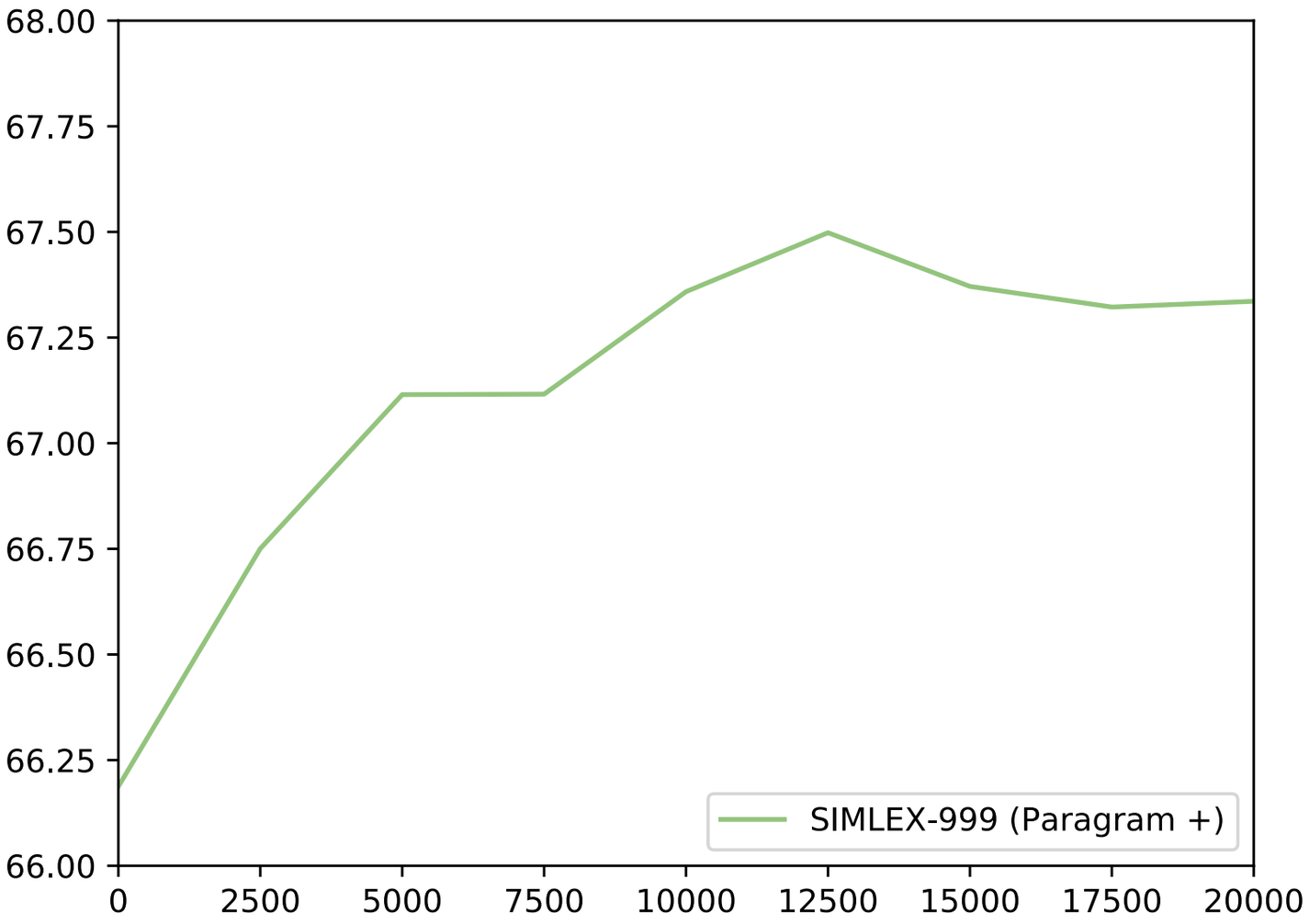
SimLex-999

Word representation evaluation



SimLex-999

Word representation evaluation

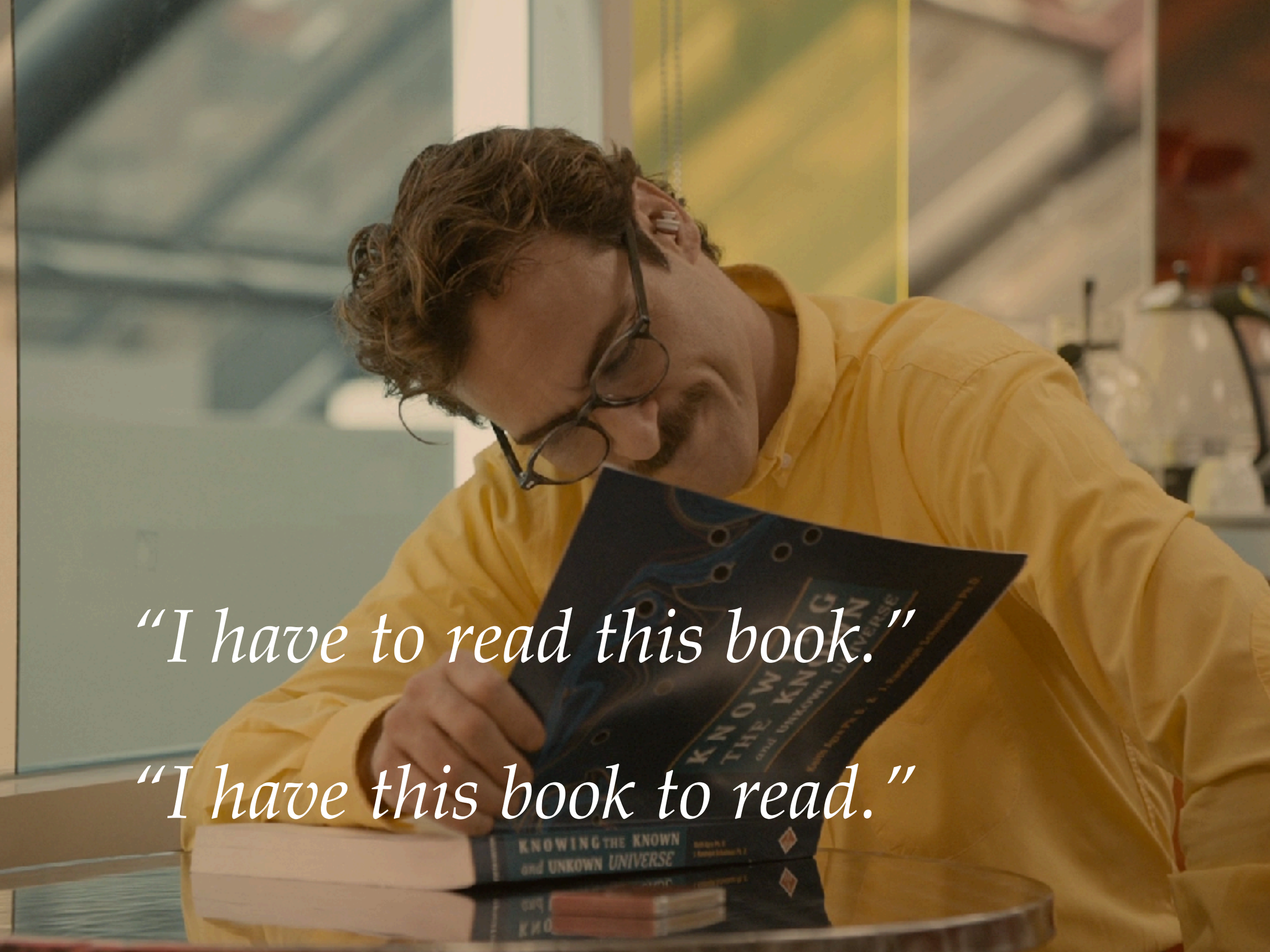


3. Neural models for composing embeddings

Instead of tuning word embeddings for algebraic composition, we now turn to *learnable composition functions*

Learnable composition functions

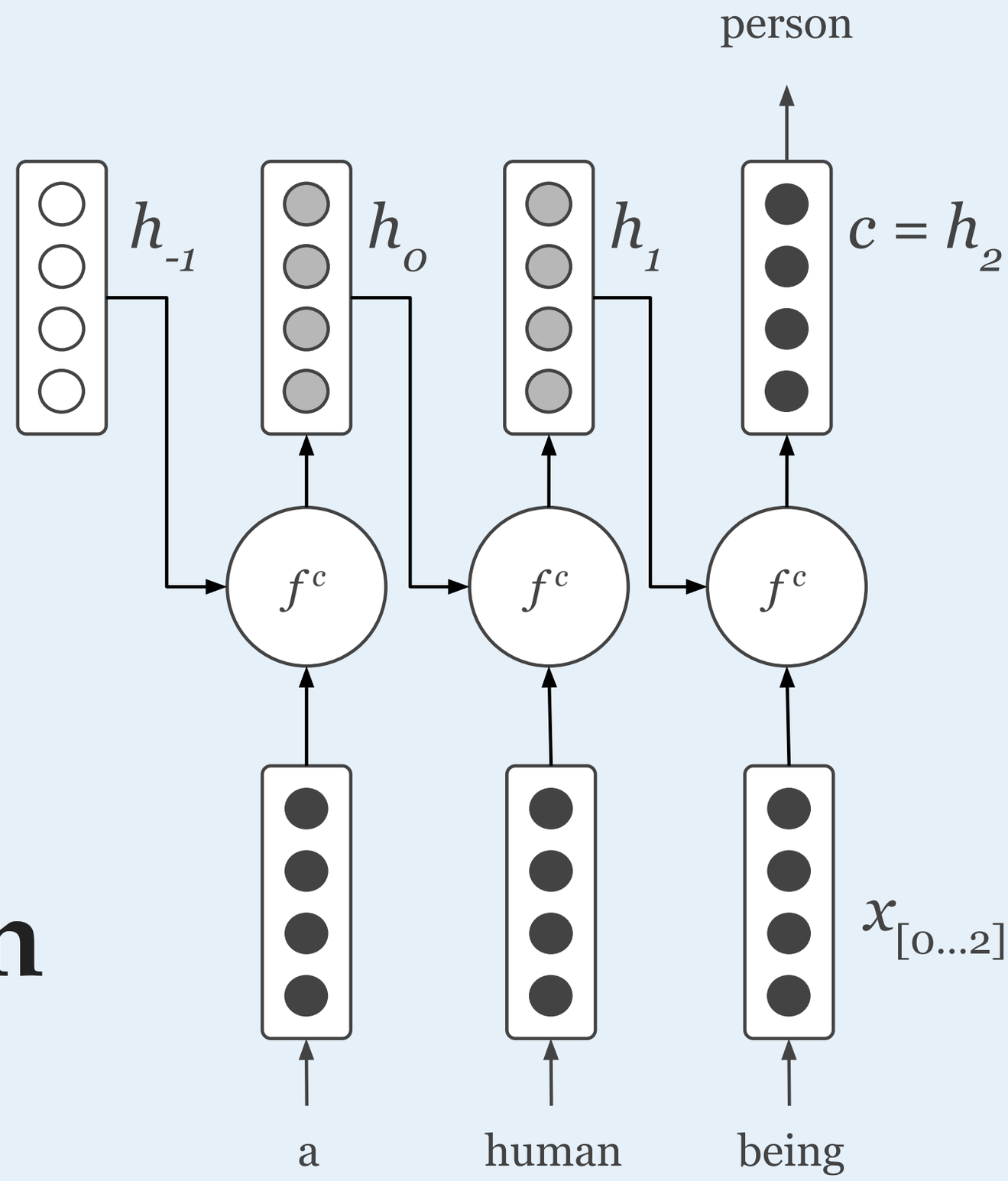
1. Projection function
2. *Recurrent composition functions*
3. Convolutional composition functions



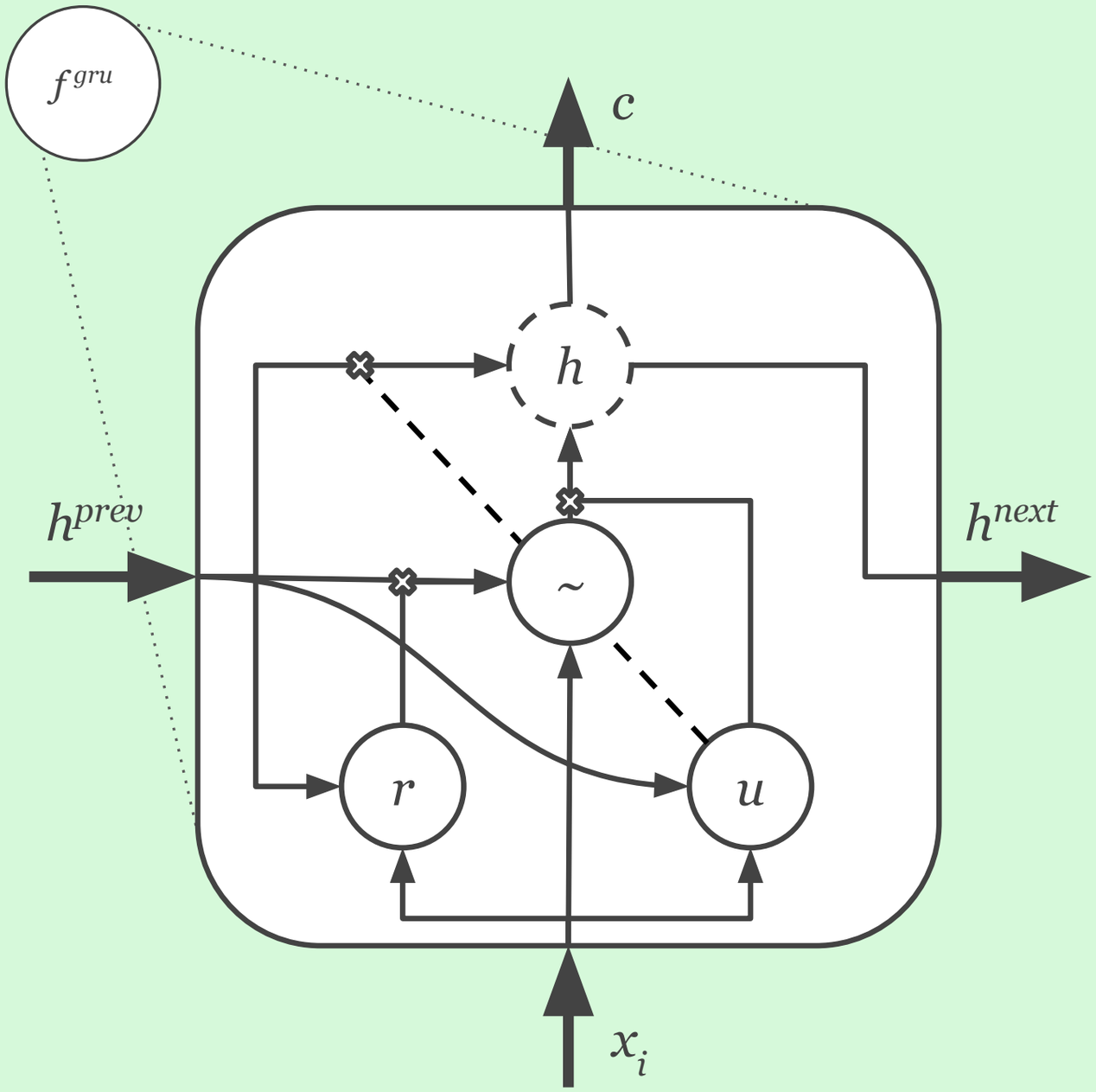
“I have to read this book.”

“I have this book to read.”

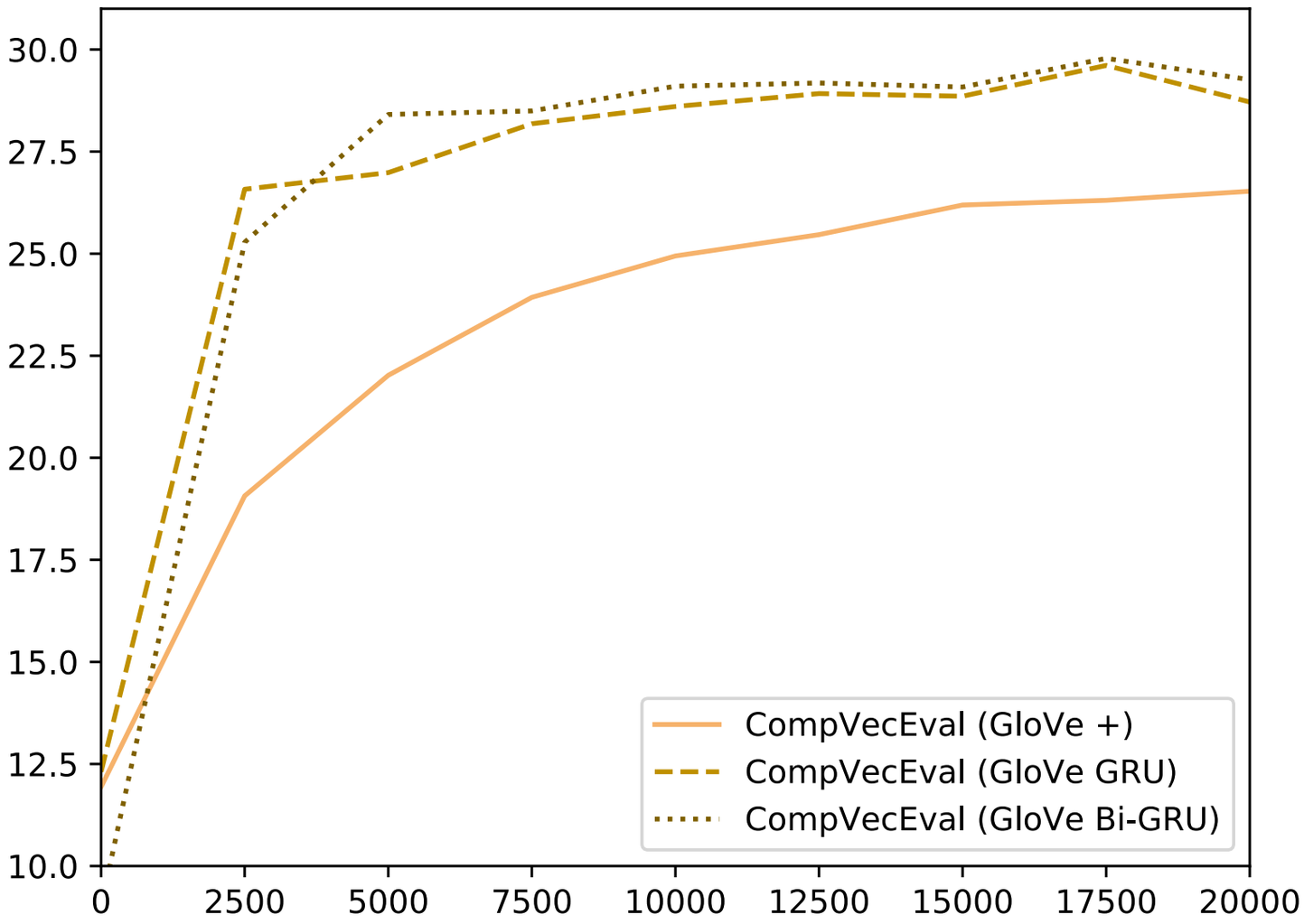
Recurrent Composition Function



Gated Recurrent Unit

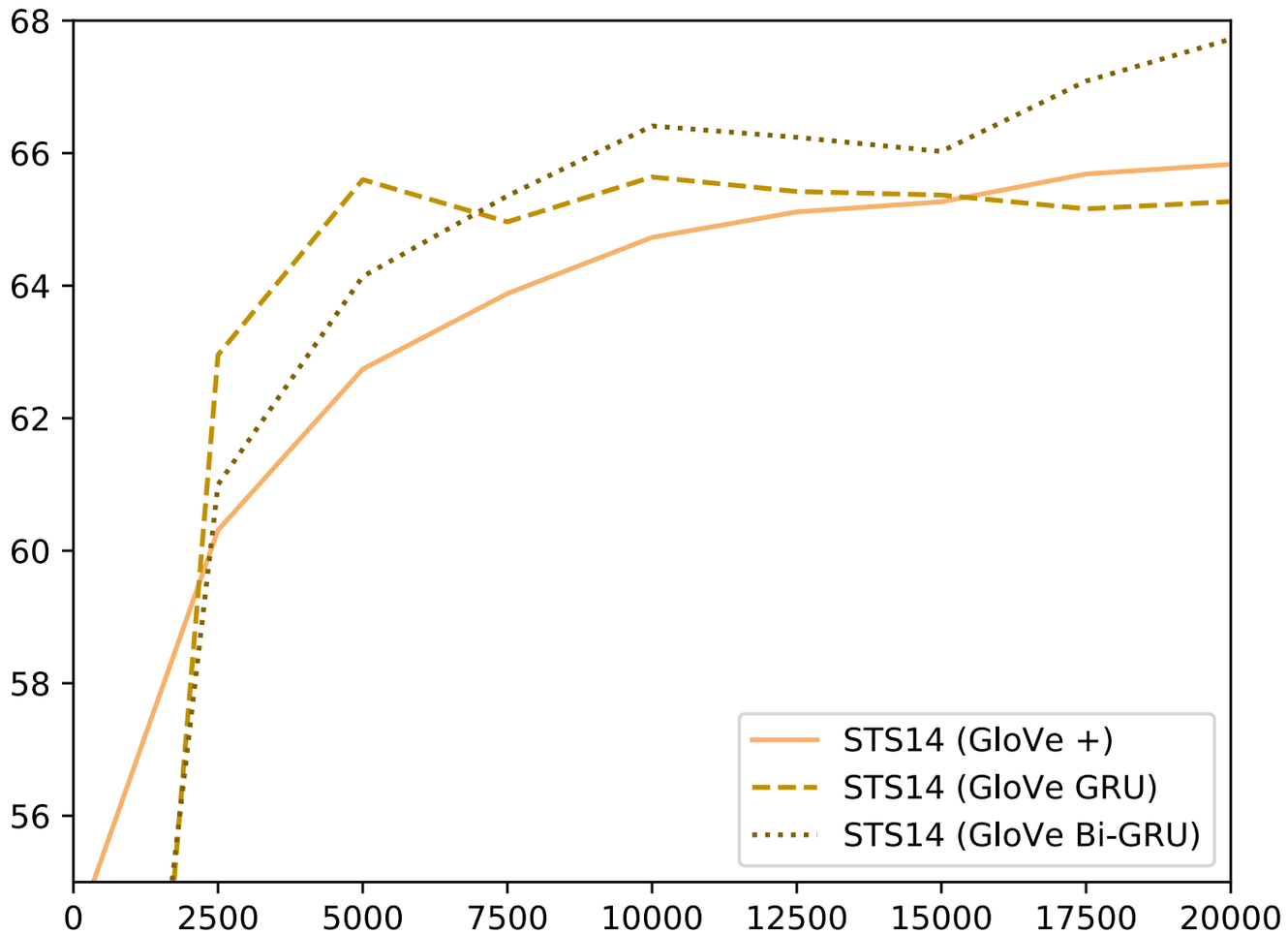


CompVecEval



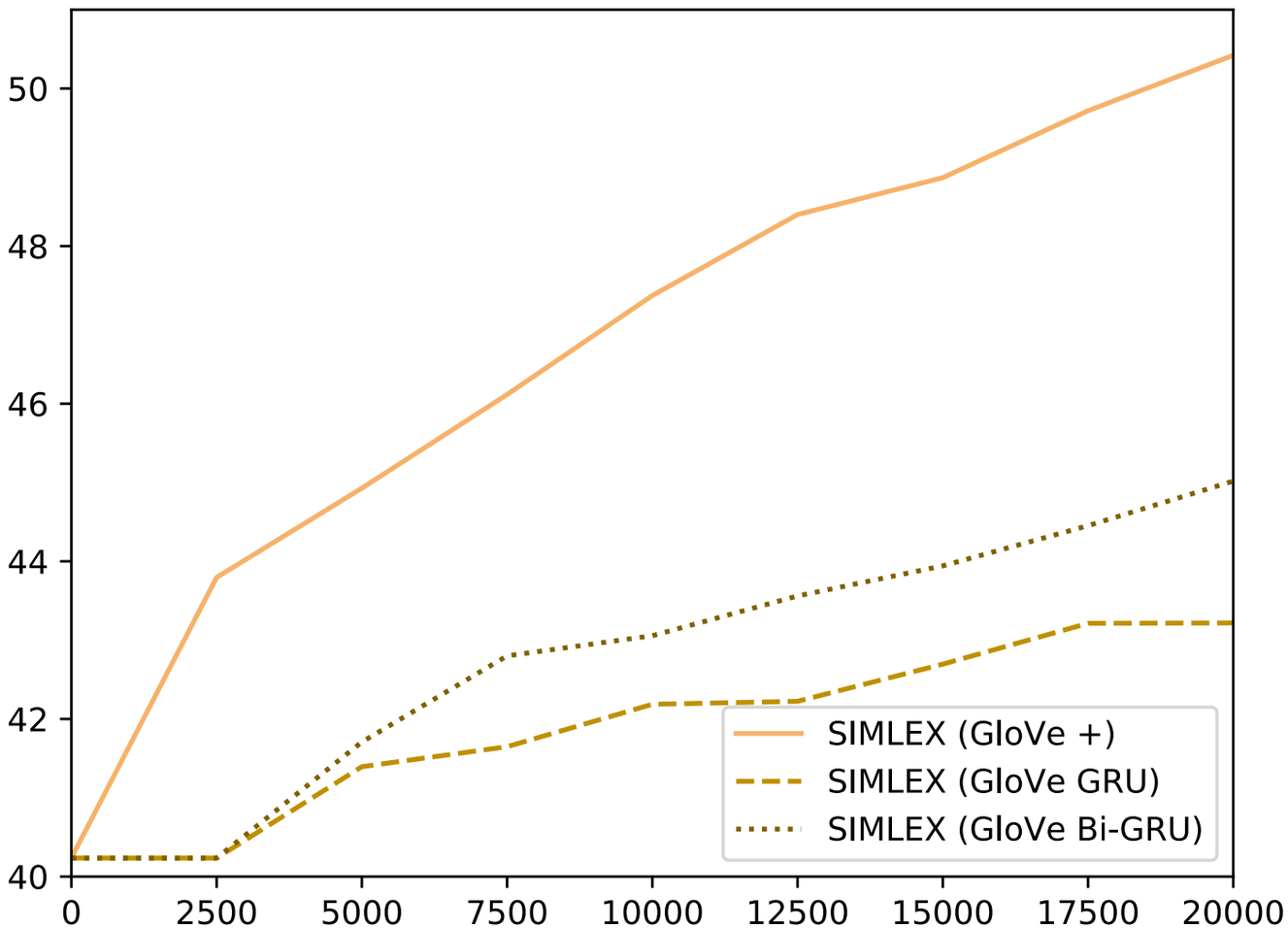
STS14 on GloVe

Sentence representation evaluation



SimLex-999 on GloVe

Word representation evaluation



Expanding to encyclopedic data

Natural language processing

From Wikipedia, the free encyclopedia

This article is about language processing by computers. For the processing of language by the human brain, see Language processing in the brain.

Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora. Challenges in natural language processing frequently involve natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, dialog systems, or some combination thereof.

Contents [hide]

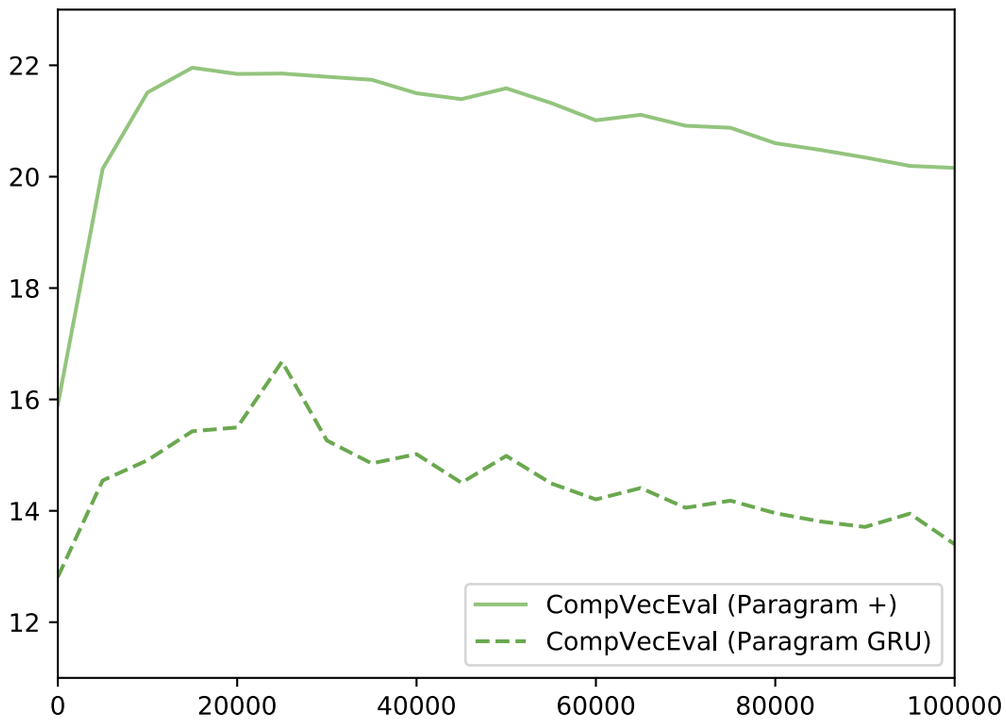
1 History
2 Statistical natural language processing (SNLP)
3 Major evaluations and tasks
3.1 Syntax
3.2 Semantics
3.3 Discourse
3.4 Speech
4 Natural language processing APIs
5 See also
6 References
7 Further reading

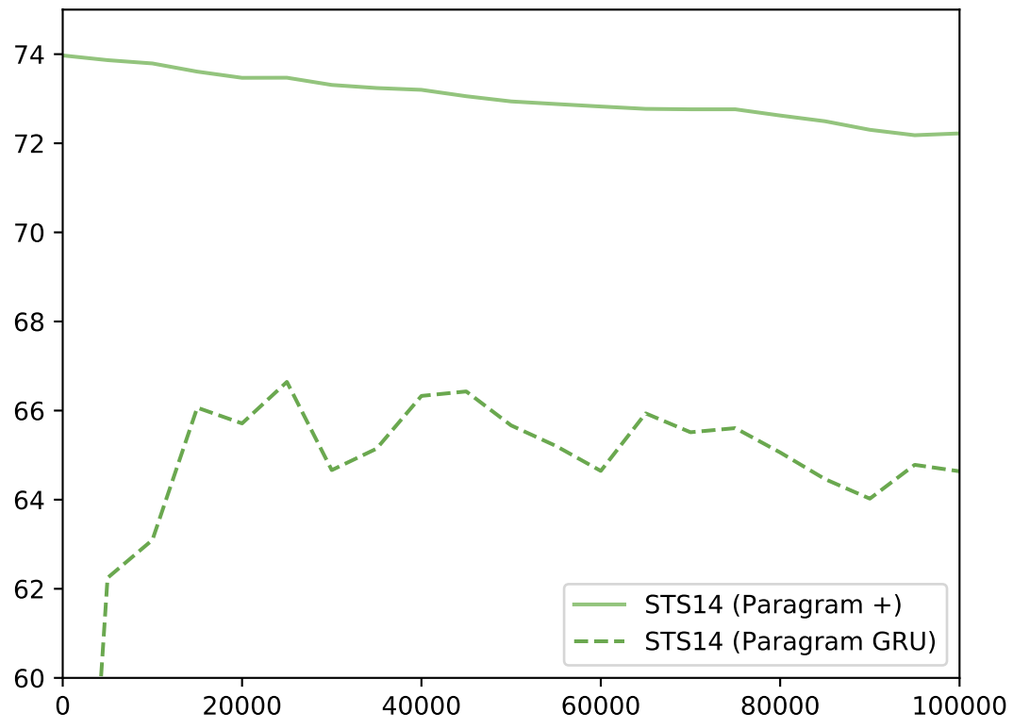
An automated online assistant providing customer service on a web page, an example of an application where natural language processing is a major component.[1]

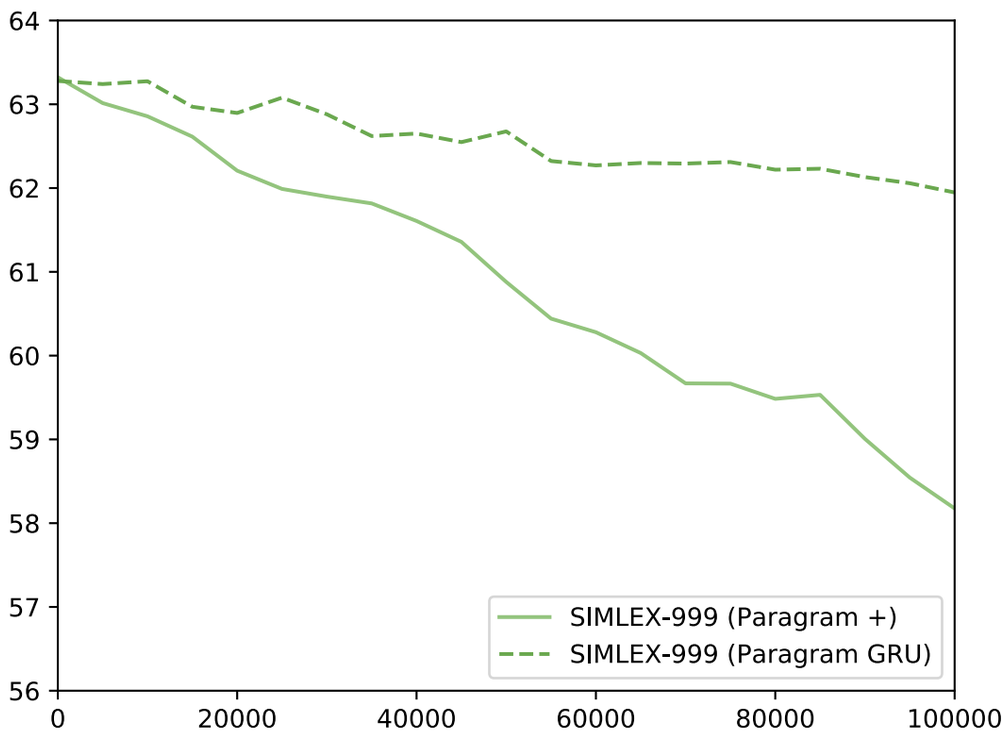
History [edit]

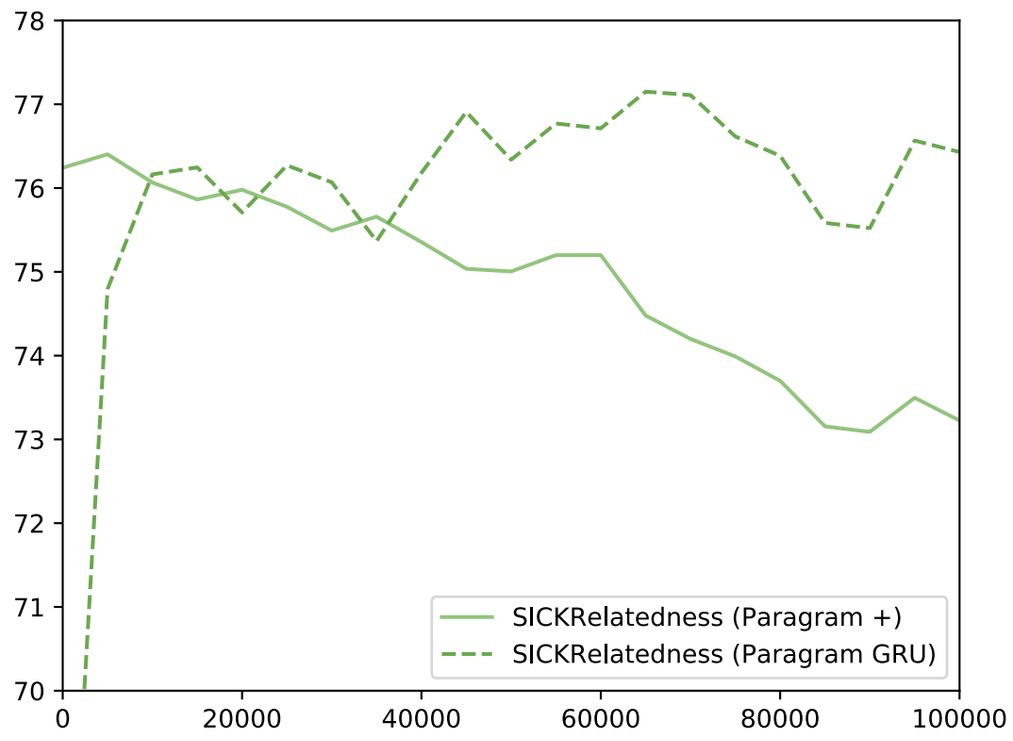
Main article: History of natural language processing

The history of NLP generally started in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.


22
20
18
16
14
12
0
20000
40000
60000
80000
100000
CompVecEval (Paragram +)
CompVecEval (Paragram GRU)


74
72
70
68
66
64
62
60
0
20000
40000
60000
80000
100000
STS14 (Paragram +)
STS14 (Paragram GRU)

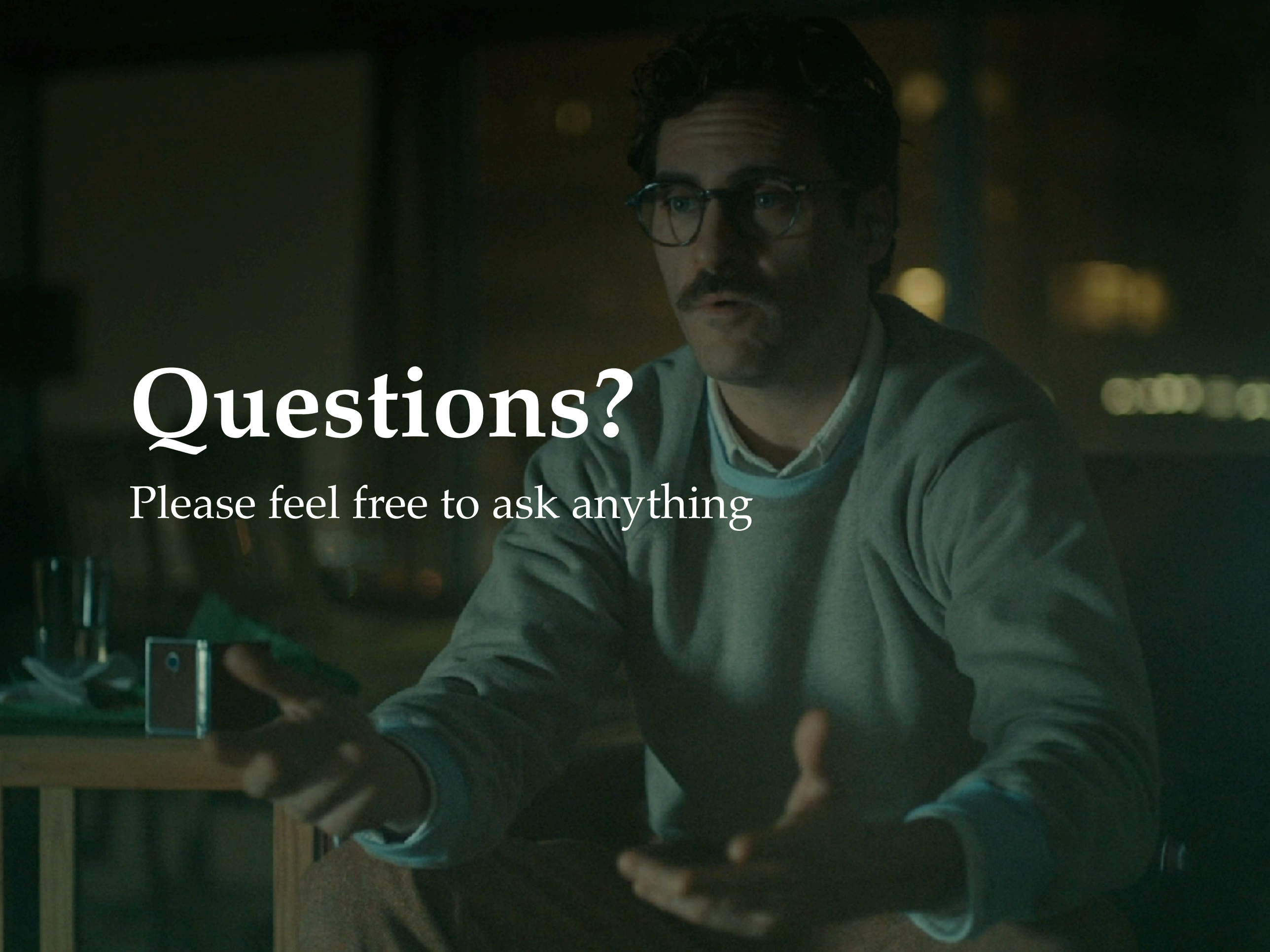

64
63
62
61
60
59
58
57
56
0
20000
40000
60000
80000
100000
SIMLEX-999 (Paragram +)
SIMLEX-999 (Paragram GRU)


78
77
76
75
74
73
72
71
70
0
20000
40000
60000
80000
100000
SICKRelatedness (Paragram +)
SICKRelatedness (Paragram GRU)

Conclusion

Semantic representations, and composition can improve when tuning using *lexicographic data*

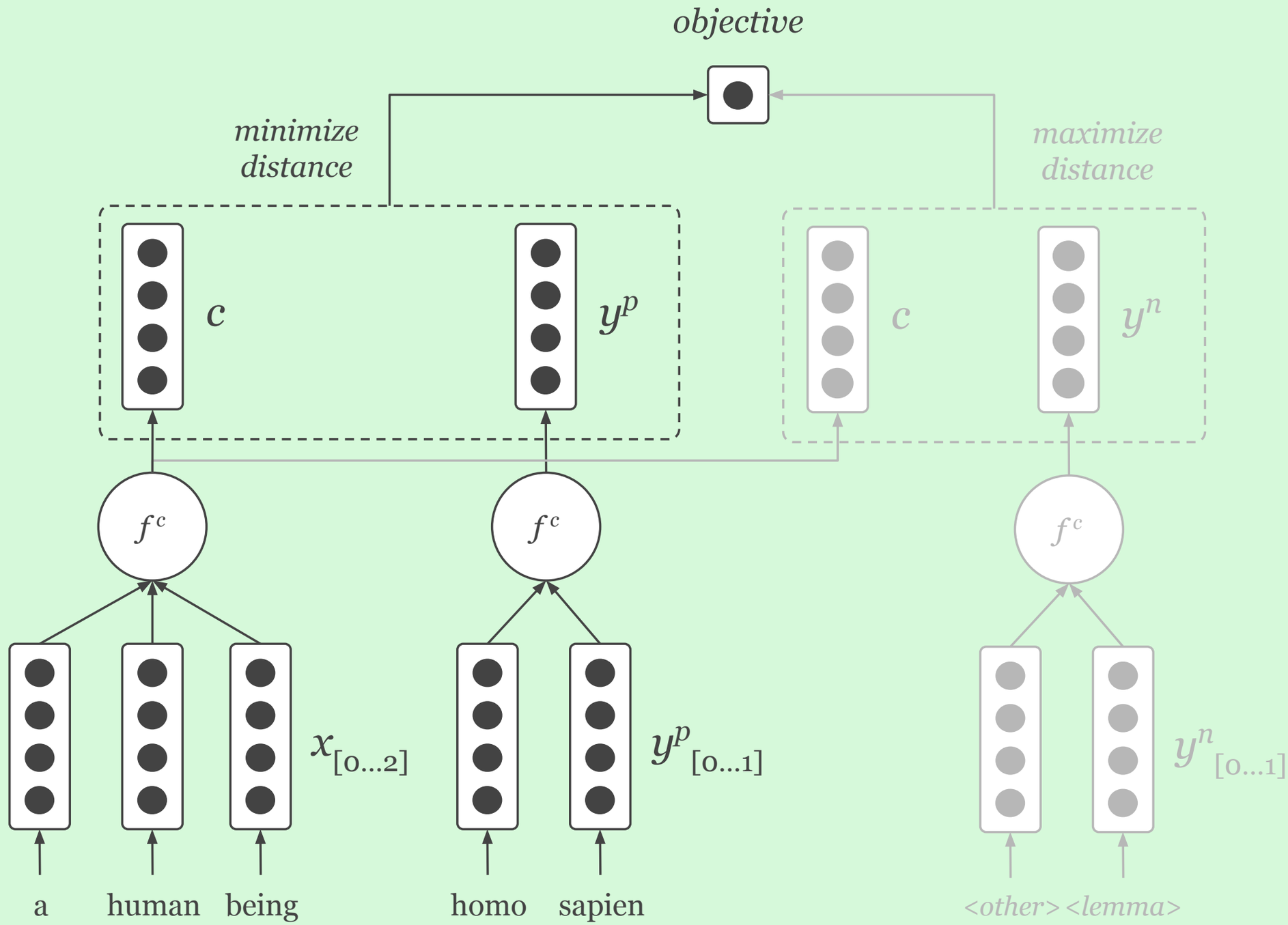
Just simple summation is a good composition function, don’t consider averaging



Questions?

Please feel free to ask anything

Multi word lemma's



CompVecEval

	Word2Vec		GloVe		fastText		Paragram	
	Org.	Tuned	Org.	Tuned	Org.	Tuned	Org.	Tuned
+	16.8	23.0 (+6.2)	11.9	26.5 (+14.5)	20.7	26.3 (+5.6)	26.5	*29.9* (+3.3)
average(d)	2.0	2.0 (-0.0)	3.3	4.1 (+0.8)	3.0	3.4 (+0.5)	3.8	*4.1* (+0.3)
+ Proj.	9.7	14.3 (+4.6)	17.5	22.7 (+5.2)	16.0	19.1 (+3.1)	20.3	*24.8* (+4.6)
RNN	8.5	7.3 (-1.2)	15.7	14.7 (-0.9)	14.2	12.6 (-1.7)	*16.3*	15.6 (-0.7)
GRU	23.4	20.7 (-2.7)	28.9	28.9 (+0.0)	27.8	26.1 (-1.6)	29.2	*29.8* (+0.6)
Bi-GRU	23.6	20.4 (-3.2)	30.2	30.1 (-0.1)	29.0	26.3 (-2.7)	29.7	**30.3** (+0.5)
CNN-3	11.4	14.8 (+3.3)	21.9	22.6 (+0.7)	22.2	22.4 (+0.2)	*24.0*	23.8 (-0.1)
CNN-3,5,7,9	12.0	14.8 (+2.8)	23.4	23.7 (+0.4)	23.3	22.6 (-0.7)	22.4	*24.2* (+1.8)

Sentence Evaluation

		Word2Vec		GloVe		fastText		Paragram	
		Org.	Tuned	Org.	Tuned	Org.	Tuned	Org.	Tuned
STS14	+	31.9	61.8 (+29.9)	54.1	65.7 (+11.5)	53.0	65.2 (+12.2)	70.5	**71.1** (+0.5)
	average(d)	32.0	41.5 (+9.6)	54.1	48.0 (-6.1)	53.2	46.3 (-6.9)	*70.5*	49.9 (-20.6)
	+ Proj.	10.1	27.1 (+16.9)	25.2	49.7 (+24.5)	22.6	39.3 (+16.8)	27.0	*55.5* (+28.5)
	RNN	41.4	46.5 (+5.2)	52.4	48.0 (-4.4)	46.8	49.8 (+3.0)	*55.7*	47.6 (-8.1)
	GRU	48.2	62.5 (+14.2)	62.9	65.5 (+2.6)	57.9	63.9 (+6.0)	65.5	*67.4* (+1.9)
	Bi-GRU	53.5	62.3 (+8.8)	66.1	67.5 (+1.4)	62.1	64.6 (+2.5)	66.8	*67.9* (+1.1)
	CNN-3	36.0	52.9 (+16.8)	55.8	59.2 (+3.4)	51.0	57.4 (+6.4)	63.6	63.5 (-0.1)
	CNN-3,5,7,9	35.5	54.4 (+18.9)	59.2	61.6 (+2.4)	54.0	59.1 (+5.1)	63.7	64.2 (+0.5)
SICK-E	+	73.8	77.3 (+3.5)	75.8	76.7 (+0.9)	76.8	76.1 (-0.7)	77.5	*78.0* (+0.6)
	average(d)	77.1	76.9 (-0.2)	79.0	78.2 (-0.8)	78.3	76.6 (-1.7)	*81.1*	78.9 (-2.1)
	+ Proj.	71.0	76.6 (+5.6)	74.0	75.4 (+1.4)	73.8	76.0 (+2.2)	75.9	*78.2* (+2.3)
	RNN	67.7	69.9 (+2.2)	69.3	68.4 (-0.9)	72.0	69.5 (-2.6)	*74.3*	67.3 (-7.0)
	GRU	75.6	78.2 (+2.6)	80.6	80.5 (-0.0)	77.7	78.7 (+1.0)	**81.5**	81.3 (-0.2)
	Bi-GRU	74.7	78.9 (+4.3)	79.7	78.7 (-1.0)	79.0	79.5 (+0.5)	81.1	*81.1* (+0.0)
	CNN-3	69.6	73.2 (+3.7)	72.9	73.8 (+0.9)	73.4	74.4 (+1.0)	74.7	*75.6* (+0.9)
	CNN-3,5,7,9	74.2	74.9 (+0.7)	74.1	*76.4* (+2.3)	75.6	75.1 (-0.6)	76.0	74.2 (-1.8)
SICK-R	+	73.4	78.4 (+5.0)	78.3	80.0 (+1.7)	78.2	78.0 (-0.2)	*80.3*	79.9 (-0.4)
	average(d)	71.4	72.2 (+0.8)	79.8	75.5 (-4.3)	79.1	74.0 (-5.0)	*81.5*	77.0 (-4.5)
	+ Proj.	66.5	70.2 (+3.7)	68.0	72.3 (+4.2)	67.0	70.4 (+3.4)	63.8	*73.4* (+9.6)
	RNN	63.8	62.4 (-1.4)	68.1	62.4 (-5.6)	68.3	63.6 (-4.8)	69.6	63.3 (-6.3)
	GRU	74.8	77.5 (+2.8)	*81.6*	81.2 (-0.4)	79.4	79.2 (-0.2)	81.3	81.1 (-0.2)
	Bi-GRU	76.9	78.3 (+1.3)	**81.8**	80.1 (-1.6)	80.1	79.1 (-1.0)	81.3	80.4 (-0.9)
	CNN-3	66.1	72.5 (+6.4)	73.4	73.3 (-0.1)	72.6	73.0 (+0.4)	*75.2*	74.4 (-0.8)
	CNN-3,5,7,9	73.9	75.6 (+1.7)	75.9	*77.9* (+2.0)	77.5	75.9 (-1.6)	*78.4*	77.4 (-0.9)

Word Evaluation

		Word2Vec		GloVe		fastText		Paragram	
		Org.	Tuned	Org.	Tuned	Org.	Tuned	Org.	Tuned
WS-353	+	70.4	70.1 (-0.3)	71.9	*76.5* (+4.6)	74.5	70.1 (-4.3)	73.1	72.1 (-1.1)
	average(d)	70.4	67.8 (-2.5)	71.9	73.2 (+1.3)	74.5	69.8 (-4.6)	73.1	**76.6** (+3.5)
	+ Proj.	70.4	70.8 (+0.4)	71.9	73.1 (+1.2)	*74.5*	72.0 (-2.4)	73.1	73.3 (+0.2)
	RNN	70.4	66.1 (-4.3)	71.9	73.2 (+1.2)	74.5	74.8 (+0.4)	73.1	*75.1* (+2.0)
	GRU	70.4	70.3 (-0.1)	71.9	72.9 (+0.9)	74.5	*76.5* (+2.1)	73.1	74.4 (+1.2)
	Bi-GRU	70.7	68.5 (-2.3)	71.9	73.2 (+1.2)	74.5	73.9 (-0.6)	73.1	*74.8* (+1.6)
	CNN-3	70.4	67.4 (-3.0)	71.9	72.5 (+0.5)	*74.5*	73.4 (-1.1)	73.1	73.6 (+0.4)
	CNN-3,5,7,9	70.4	68.6 (-1.7)	71.9	71.8 (-0.1)	*74.5*	72.7 (-1.7)	73.1	72.6 (-0.5)
SimLex	+	44.0	51.4 (+7.4)	40.2	50.1 (+9.8)	37.3	46.5 (+9.1)	66.2	*67.0* (+0.8)
	average(d)	44.0	37.5 (-6.5)	40.2	45.0 (+4.8)	37.3	35.5 (-1.8)	*66.2*	62.1 (-4.1)
	+ Proj.	44.0	**46.6** (+2.6)	40.2	44.4 (+4.2)	37.3	43.6 (+6.3)	**66.2**	*66.7* (+0.6)
	RNN	44.0	44.3 (+0.3)	40.2	43.5 (+3.2)	37.3	39.7 (+2.4)	*66.2*	66.1 (-0.1)
	GRU	44.0	47.3 (+3.3)	40.2	44.6 (+4.3)	37.3	43.8 (+6.4)	66.2	**67.2** (+1.0)
	Bi-GRU	44.0	51.1 (+7.1)	40.2	**45.4** (+5.2)	37.3	43.3 (+6.0)	66.2	*67.2* (+1.0)
	CNN-3	44.0	**46.2** (+2.2)	40.2	44.0 (+3.7)	37.3	41.6 (+4.3)	66.2	*66.9* (+0.7)
	CNN-3,5,7,9	44.0	47.0 (+3.0)	40.2	44.3 (+4.1)	37.3	40.3 (+2.9)	**66.2**	*67.2* (+1.0)
SimVerb	+	34.2	39.3 (+5.2)	25.4	34.2 (+8.7)	23.0	34.1 (+11.1)	56.8	**58.3** (+1.5)
	average(d)	34.2	30.7 (-3.4)	25.4	32.2 (+6.8)	23.0	29.2 (+6.2)	*56.8*	55.7 (-1.1)
	+ Proj.	34.2	35.7 (+1.5)	25.4	28.5 (+3.0)	23.0	29.4 (+6.4)	56.8	*57.3* (+0.5)
	RNN	34.2	29.4 (-4.8)	25.4	27.1 (+1.7)	23.0	24.3 (+1.3)	*56.8*	56.0 (-0.8)
	GRU	34.2	37.2 (+3.1)	25.4	29.7 (+4.2)	23.0	29.5 (+6.5)	56.8	*57.9* (+1.1)
	Bi-GRU	34.2	36.9 (+2.8)	25.4	29.5 (+4.1)	23.0	29.9 (+6.9)	56.8	*58.1* (+1.3)
	CNN-3	34.2	33.1 (-1.1)	25.4	28.3 (+2.9)	23.0	26.2 (+3.2)	56.8	*57.6* (+0.8)
	CNN-3,5,7,9	34.2	33.0 (-1.1)	25.4	27.8 (+2.4)	23.0	26.6 (+3.6)	56.8	*57.3* (+0.5)

Vocabulary Overlap

CompVecEval

	WordNet		Wikipedia	
Total vocabulary size	48,944		378,950	
word2vec	34,976	71.4 %	83,888	22.1 %
GloVe	42,742	87.3 %	181,967	48.0 %
fastText	41,554	84.9 %	320,708	85.7 %
Paragram	44,138	90.2 %	297,440	78.5 %

CompVecEval

Alternative Vocabulary

		Word2Vec	fastText	GloVe
+	MNR	0.094	**0.075**	0.094
	MRR	4.983 %	6.852 %	2.734 %
	MAP	4.304 %	6.153 %	2.202 %
	MP@10	1.617 %	1.947 %	8.167 %
×	MNR	0.444	0.439	*0.342*
	MRR	0.107 %	0.196 %	0.154 %
	MAP	0.090 %	0.182 %	0.133 %
	MP@10	0.017 %	0.037 %	0.023 %
average(d)	MNR	0.287	0.210	*0.168*
	MRR	1.358 %	2.309 %	1.033 %
	MAP	1.143 %	2.020 %	0.798 %
	MP@10	0.270 %	0.530 %	0.243 %
max(d)	MNR	0.297	0.222	*0.180*
	MRR	0.841 %	2.065 %	0.796 %
	MAP	0.688 %	1.800 %	0.645 %
	MP@10	0.187 %	0.437 %	0.183 %
LSTM	MNR	0.376	0.309	*0.234*
	MRR	0.074 %	0.095 %	0.118 %
	MAP	0.069 %	0.084 %	0.098 %
	MP@10	0.017 %	0.006 %	0.020 %
Random	MNR	0.448		
	MRR	0.027 %		
	MAP	0.022 %		
	MP@10	0.003 %		

Vocabulary Overlap

SENTENCE REPRESENTATION EVALUATION

		word2Vec					GloVe					fastText					Paragram				
	Task	Emb.	WN-e	WN-r	Wiki-e	Wiki-r	Emb.	WN-e	WN-r	Wiki-e	Wiki-r	Emb.	WN-e	WN-r	Wiki-e	Wiki-r	Emb.	WN-e	WN-r	Wiki-e	Wiki-r
MR	20303	15965	11178	385	14621	2019	18589	11559	4	16323	317	17369	11256	307	15665	975	17748	11375	188	15811	829
CR	5674	4965	4042	121	4729	316	5472	4161	2	5026	19	4998	4035	128	4723	322	4763	4038	125	4578	467
SUBJ	22616	17515	12327	528	16465	2937	20855	12850	5	18990	412	19872	12499	356	18297	1105	17629	12140	715	16545	2857
MPQA	6238	6083	5114	50	5802	107	6197	5162	2	5898	11	6179	5141	23	5877	32	5676	4974	190	5506	403
SST	17558	14388	10370	357	13241	1641	16525	10724	3	14687	195	15373	10419	308	14059	823	16830	10726	1	14880	2
TREC	8968	7395	6124	462	7170	1186	8599	6575	11	8252	104	8244	6313	273	7890	466	7009	5937	649	6853	1503
MRPC	17908	12203	9343	571	11876	3563	16074	9907	7	15001	438	14233	9477	437	13488	1951	13292	9467	447	12862	2577
SICK-E	2312	2277	2023	14	2152	13	2305	2035	2	2165	0	2291	2030	7	2157	8	2312	2035	2	2165	0
SICK-R	2312	2277	2023	14	2152	13	2305	2035	2	2165	0	2291	2030	7	2157	8	2312	2035	2	2165	0
STS-B	16012	12093	9438	474	11718	2419	14634	9905	7	13802	335	13477	9564	348	12827	1310	16012	9762	150	13390	747
STS12	8124	6716	5907	240	6496	939	7766	6145	2	7352	83	7147	5932	215	6833	602	7867	6142	5	7403	32
STS13	5191	4499	4047	128	4350	352	4910	4173	2	4667	35	4678	4069	106	4489	213	4896	4170	5	4674	28
STS14	9268	7748	6491	218	7488	781	8586	6707	2	8136	133	8245	6546	163	7842	427	8682	6704	5	8192	77
STS15	7431	6410	5425	189	6169	572	7047	5611	3	6665	76	6706	5461	153	6372	369	7097	5607	7	6684	57
STS16	3988	3489	3124	101	3352	291	3804	3223	2	3605	38	3689	3149	76	3502	141	3559	3162	63	3423	220

Statistical Significance

GloVe and Word2Vec

	Word2Vec +							GloVe +						
	n	μ_{pre}	σ_{pre}	μ_{tuned}	σ_{tuned}	p	t	n	μ_{pre}	σ_{pre}	μ_{tuned}	σ_{tuned}	p	t
CompVecEval (MRR)	12	16.8	0.43	23.0	0.29	4.5e-22	-40.1	12	11.9	0.01	26.6	0.14	3.0e-42	-333.5
MNR	12	83.9	0.59	87.9	0.12	1.7e-16	-22.1	12	83.5	0.01	92.2	0.07	2.6e-44	-414.0
MAP	12	15.3	0.39	21.2	0.23	9.3e-23	-43.2	12	10.8	0.01	24.6	0.14	8.5e-42	-318.2
MP@10	12	2.9	0.06	4.4	0.05	4.1e-26	-61.5	12	2.2	0.00	4.9	0.03	6.6e-40	-261.2
WS-353	12	70.4	0.09	70.6	1.25	**7.5e-01**	**-0.3**	12	71.9	0.00	76.2	0.50	9.4e-19	-28.2
WS-353 (Sim.)	12	80.3	0.01	79.6	1.08	**4.0e-02**	**2.2**	12	81.1	0.00	82.6	0.61	3.6e-08	-8.2
WS-353 (Rel.)	12	63.5	0.18	60.8	2.05	2.3e-04	4.4	12	65.9	0.00	71.2	0.94	4.4e-15	-18.9
MC-30	12	83.8	0.00	84.8	2.97	**2.7e-01**	**-1.1**	12	82.6	0.00	86.2	2.88	4.3e-04	-4.1
RG-65	12	76.1	0.00	80.2	1.81	1.4e-07	-7.6	12	76.8	0.00	83.0	1.34	2.9e-13	-15.4
Rare Word	12	58.0	0.11	55.1	0.85	1.5e-10	11.2	12	52.5	0.03	57.1	0.49	1.5e-19	-30.7
MTurk-3k	12	72.9	0.24	69.3	0.48	9.2e-17	22.7	12	80.0	0.00	79.0	0.31	2.9e-10	10.8
MTurk-287	12	66.5	0.77	66.6	1.86	**9.1e-01**	**-0.1**	12	72.3	0.00	73.5	0.53	3.0e-07	-7.2
MTurk-771	12	66.8	0.00	69.3	0.78	3.3e-10	-10.7	12	70.9	0.00	72.2	0.50	1.2e-08	-8.8
YP-130	12	49.8	0.00	64.0	1.88	1.2e-17	-25.0	12	51.9	0.00	60.8	1.03	5.1e-19	-29.0
SimLex	12	44.0	0.00	51.2	0.64	2.0e-21	-37.5	12	40.2	0.00	50.4	0.54	3.2e-26	-62.2
Verb-143	12	49.7	0.00	36.0	3.35	3.5e-12	13.6	12	35.8	0.00	29.1	1.28	2.5e-14	17.4
SimVerb	12	34.2	0.00	40.0	0.69	1.2e-18	-27.9	12	25.4	0.00	34.6	0.28	2.0e-31	-107.5
MR	12	75.3	0.15	74.3	0.29	1.0e-09	10.1	12	76.2	0.05	76.0	0.20	**2.1e-02**	**2.5**
CR	12	77.2	0.29	77.0	0.53	**3.3e-01**	**1.0**	12	78.9	0.07	78.7	0.33	**1.0e-02**	**2.8**
SUBJ	12	89.6	0.15	88.9	0.27	7.1e-08	7.9	12	90.4	0.03	90.0	0.18	6.6e-07	6.9
MPQA	12	86.2	0.11	85.3	0.20	1.3e-12	14.3	12	86.0	0.00	85.8	0.19	**3.1e-02**	**2.3**
SST	12	79.3	0.36	78.1	0.45	4.2e-07	7.1	12	79.4	0.08	78.7	0.42	3.1e-05	5.2
TREC	12	75.1	1.55	78.2	1.06	1.7e-05	-5.5	12	81.2	0.27	80.1	1.60	**3.5e-02**	**2.3**
MRPC	12	68.7	0.80	69.3	0.86	**1.3e-01**	**-1.6**	12	69.6	0.74	69.7	0.80	**9.4e-02**	**1.7**
SICK-E	12	73.8	0.63	77.0	0.49	7.5e-12	-13.1	12	75.8	0.00	76.8	0.71	2.2e-04	-4.4
SICK-R	12	73.4	0.73	78.2	0.53	1.8e-14	-17.6	12	78.3	0.00	79.8	0.29	3.0e-14	-17.2
STS14	12	31.9	0.23	62.1	0.25	5.9e-41	-291.5	12	54.1	0.00	65.7	0.11	4.4e-43	-364.0
STS15	12	36.9	0.49	68.6	0.35	5.6e-36	-173.1	12	58.1	0.00	66.6	0.14	3.4e-37	-196.7

Statistical Significance

fastText and Paragram

	fastText +							Paragram +						
	n	μ_{pre}	σ_{pre}	μ_{tuned}	σ_{tuned}	p	t	n	μ_{pre}	σ_{pre}	μ_{tuned}	σ_{tuned}	p	t
CompVecEval (MRR)	12	20.7	0.04	26.6	0.20	1.4e-30	-98.2	17	26.5	0.02	29.8	0.10	2.9e-44	-121.9
MNR	12	86.3	0.03	90.9	0.07	3.1e-37	-197.5	17	90.3	0.02	92.8	0.04	4.0e-51	-199.8
MAP	12	18.9	0.04	24.6	0.17	1.3e-31	-109.5	17	24.8	0.02	27.8	0.10	2.5e-44	-122.4
MP@10	12	3.6	0.01	4.8	0.05	1.7e-28	-79.0	17	5.2	0.00	5.7	0.05	6.5e-32	-49.9
WS-353	12	74.5	0.00	69.7	0.86	8.7e-15	18.3	17	73.1	0.00	70.3	1.01	1.5e-12	11.1
WS-353 (Sim.)	12	78.7	0.00	73.8	1.02	1.3e-13	16.0	17	80.6	0.00	78.5	1.03	1.6e-09	8.3
WS-353 (Rel.)	12	69.5	0.00	61.8	1.21	4.6e-16	21.0	17	67.0	0.00	62.4	1.15	1.1e-16	15.8
MC-30	12	85.5	0.00	76.6	1.95	4.1e-13	15.1	*17*	*74.1*	*0.00*	*76.9*	*3.21*	**1.4e-03**	**-3.5**
RG-65	12	79.4	0.00	76.2	1.71	4.0e-06	6.1	17	73.6	0.00	77.8	1.29	2.1e-14	-13.1
Rare Word	12	54.3	0.00	53.7	0.28	3.9e-06	6.1	17	63.5	0.05	63.8	0.29	7.8e-04	-3.7
MTurk-3k	12	75.8	0.00	68.6	0.55	8.9e-23	43.2	17	75.9	0.00	74.6	0.23	6.9e-21	22.0
MTurk-287	12	68.2	0.00	61.7	1.39	2.5e-13	15.5	17	67.0	0.00	64.7	0.72	9.4e-14	12.4
MTurk-771	12	66.9	0.00	66.1	0.70	7.8e-04	3.9	17	70.1	0.00	71.5	0.40	1.8e-14	-13.2
YP-130	12	49.9	0.00	58.2	2.12	7.6e-12	-13.1	17	65.5	0.00	68.6	1.17	5.8e-12	-10.6
SimLex	12	37.3	0.00	46.1	0.53	4.7e-25	-55.0	17	66.2	0.00	67.2	0.28	1.7e-15	-14.4
Verb-143	12	38.9	0.00	29.7	3.51	1.4e-08	8.7	17	58.2	0.00	54.4	1.44	5.0e-12	10.6
SimVerb	12	23.0	0.00	33.2	0.81	1.8e-22	-41.9	17	56.8	0.00	57.9	0.27	5.4e-17	-16.2
MR	12	74.0	0.20	73.5	0.20	1.8e-05	5.5	17	74.2	0.00	74.5	0.25	1.9e-06	-5.8
CR	*12*	*76.9*	*0.32*	*76.9*	*0.59*	**9.4e-01**	**0.1**	*17*	*78.0*	*0.07*	*77.8*	*0.39*	**5.7e-02**	**2.0**
SUBJ	12	89.8	0.16	89.1	0.19	1.5e-09	9.9	17	88.1	0.01	88.3	0.19	8.3e-04	-3.7
MPQA	12	85.6	0.08	85.0	0.26	7.4e-07	6.8	*17*	*86.0*	*0.00*	*86.0*	*0.17*	**2.1e-01**	**1.3**
SST	*12*	*77.2*	*0.35*	*77.2*	*0.60*	**8.6e-01**	**0.2**	17	79.1	0.00	78.7	0.33	7.1e-05	4.6
TREC	*12*	*79.5*	*0.87*	*78.8*	*1.17*	**1.1e-01**	**1.7**	*17*	*81.0*	*0.11*	*80.9*	*0.78*	**4.4e-01**	**0.8**
MRPC	*12*	*68.4*	*0.66*	*68.5*	*0.72*	**5.8e-01**	**-0.6**	17	69.9	0.00	70.9	0.69	1.4e-06	-5.9
SICK-E	*12*	*76.8*	*0.41*	*76.7*	*0.46*	**6.0e-01**	**0.5**	17	77.5	0.00	77.8	0.22	1.2e-07	-6.8
SICK-R	*12*	*78.2*	*0.12*	*78.1*	*0.90*	**6.1e-01**	**0.5**	*17*	*80.3*	*0.00*	*80.2*	*0.20*	**4.7e-03**	**3.0**
STS14	12	53.0	0.13	65.1	0.23	6.6e-35	-154.7	17	70.5	0.00	71.3	0.13	4.2e-21	-22.4
STS15	12	58.1	0.17	68.6	0.23	9.0e-33	-123.7	17	75.0	0.00	75.1	0.09	2.0e-07	-6.6

Improving the Compositionality of Word Embeddings

Master Thesis

Author:
Thijs Scheepers

Supervisors:
dr. Evangelos Kanoulas
dr. Efstratios Gavves

Thanks!

Announcements

UNIVERSITEIT VAN AMSTERDAM

MASTER THESIS

Improving the Compositionality of Word Embeddings

Author:
Mathijs J. SCHEEPERS

Supervisors:
dr. Evangelos KANOULAS
dr. Efstratios GAVVES

Assessor:
prof.dr. Maarten DE RIJKE

A thesis submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of ***Master of Science*** *in* ***Artificial Intelligence****.*

November 29, 2017

v

3.3.3 V-
overlap with the pretrained embeddings 26
racticalities 26
Evaluation 27
or Ranking 27
es 28
cEval 29
. 30
. 30
. 32
. 33
braic composition **35**
nality 36
. 36
. 37
. 38
targets 39
quality 40
cient 40
ficient 40
. 41
. 42
. 42
. 43
. 43
. 43
. 44
anical Turk 44
. 45
. 46
. 46
ation 47
. 48
. 49
. 50
. 51
. 53
. 53
. 57
. 59
. 59
. 61
63
. 64

ults from evaluating ranking. All results for MRR, are denoted as ×100 in percentages. We also in- random rankings on the dataset, this is not specific to any of the word embeddings.

	Word2Vec	GloVe	fastText	Paragram
om	0.7 %			
	16.8 %	11.9 %	20.7 %	**26.5 %**
d)	2.0 %	3.3 %	3.0 %	*3.8 %*
	0.6 %	0.9 %	0.9 %	*1.0 %*
(d)	6.6 %	13.7 %	14.6 %	*20.5 %*
dom	54.2 %			
	83.9 %	83.5 %	86.3 %	**90.3 %**
$g(d)$	71.7 %	75.5 %	71.2 %	71.2 %
	62.8 %	65.2 %	59.0 %	54.6 %
$ax(d)$	63.1 %	83.7 %	78.5 %	*85.7 %*
ndom	0.6 %			
	15.3 %	10.8 %	18.9 %	**24.8 %**
$vg(d)$	1.8 %	2.9 %	2.6 %	*3.4 %*
×	0.6 %	0.8 %	0.8 %	*0.9 %*
$max(d)$	6.0 %	12.4 %	13.3 %	*18.9 %*
random	0.1 %			
+	2.9 %	2.2 %	3.6 %	**5.2 %**
$avg(d)$	0.3 %	0.6 %	0.5 %	*0.8 %*
×	0.1 %	0.1 %	0.1 %	*0.3 %*
$max(d)$	1.1 %	2.4 %	2.4 %	*3.8 %*

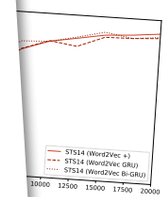
pretrained embeddings, Paragram is the clear winner. It is also the which is already tuned from an original embedding (namely GloVe). per by [135] they already showed significant improvements over e evaluation tasks, so this is in line with our results here.

ge margin between the performance of composing by averaging and sition. Clearly averaging is not necessarily the best approach when eddings using not just their angle. We also see that multiplicative early fails to amount to a meaningful representation as the results are to the random baseline. This is in direct contradiction of the state- y Mitchell and Lapata [89].

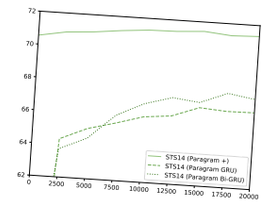
by max-pooling also seems to work surprisingly well. Especially, when he loss of data inherit in the operation. The operation will discretely aximum value on the embedding dimension. Clearly this seems to have mantic value.

erforms relatively poorly on this compositional dataset, but this can be explained by the nature of the test data. The dataset where Word2Vec was trained on used news data, where fastText and GloVe use more definitional data, Wikipedia and Common Crawl respectively. Still lots of researchers use Word2Vec as a starting point for their training procedure, our evaluation here shows that there are better options.

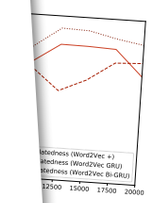
fastText is able to perform well on additive composition. When comparing it to GloVe,



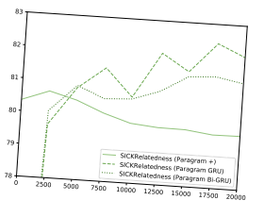
Word2Vec



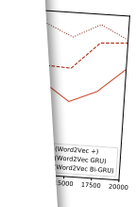
(B) STS14 Paragram



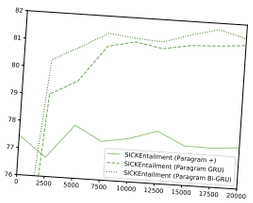
s Word2Vec



(D) SICK Relatedness Paragram



rd2Vec



(F) SICK Entailment Paragram

ure shows the progression of sentence represen- res during tuning of word2vec, and Paragram nd Bi-GRU composition. We display Pearsons's and SICK-R and accuracy ×100 for SICK-E.

word embed- thods in Natural r 17-21, 2015. Ed. Linguistics, 2015, org/anthology/D/

rent neural networks". . 2673–2681. l information processing

n act theory and pragmatics

e theory of speech acts. Cam-

shelf: an astounding baseline ference on computer vision and

onal Vector Grammars." In: ACL.

els for semantic compositionalit of the conference on empirical metho l. 1631. Citeseer. 2013, p. 1642.

sitionality through recursive mat joint conference on empirical metho tional natural language learning. Ass 012, pp. 1201–1211.

measurement of association betwe sychology 15.1 (1904), pp. 72–101.

a simple way to prevent neural n achine learning research 15.1 (2014),

eston, Rob Fergus, et al. "End-to-en al information processing systems. 201

, and Quoc V Le. "Sequence to seque Advances in neural information proce

ompositionality". In: *Stanford encyclo*

Szabó<. "Compositionality". In: *The St* Edward N. Zalta. Summer 2017. Meta ity, 2017.

http://github.com/tscheepers/CompVec

LABEL385

LABEL385

Enschede
Rightersbleek-Zandvoort 10 2.06
7521 BE Enschede
+31 (0)53 711 34 99

LABEL385

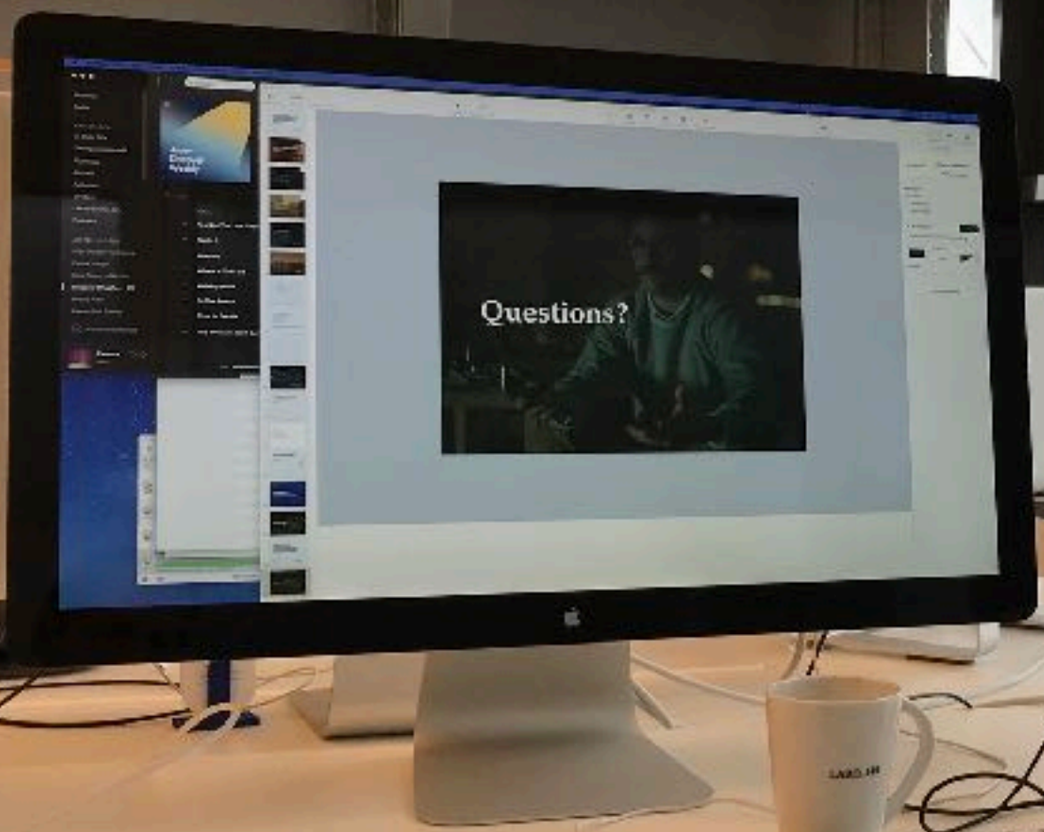
Enschede
Rightersbleek-Zandvoort 10 2.06
7521 BE Enschede
+31 (0)53 711 34 99

Amsterdam
Kruithuisstraat 13
1018 WJ Amsterdam
+31 (0)20 261 47 49


LABEL305
LABEL305
Amsterdam
Kruithuisstraat 13
1018 WJ Amsterdam
+31 (0)20 261 47 49
Noorderkerk
Sexmuseum
Frank Huis
Koninklijk Paleis Amsterdam
Abraxas
DE WALLEN
CENTRUM
NIEUWMARKT EN LASTAGE
Amsterdam Museum
Museum Het Rembrandthuis
Allard Pierson Museum
DoubleTree by Hilton Amsterdam
Muziekgebouw aan 't IJ
Java-eiland
NEMO Science
Club Panama Amsterdam
Winkelcentrum Brazilië
RIETLANDEN
ZEEBURG
CZAAR PETERBUURT
Verzetsmuseum
Bloemenmarkt
Pathé Tuschinski
Hortus Botanicus Amsterdam
ARTIS
De Gooyer
OOSTELIJK HAVENGEBIED
Hermitage Amsterdam
Pathé City
Foam
Magere Brug
WEESPERBUURT EN PLANTAGE
Koninklijk Theater Carré
Tropenmuseum
INDISCHE BUURT
Dappermarkt
Oosterpark
Rijksmuseum
Heineken Experience
OLVC, locatie Oost
Albert Cuypmarkt
OUDE PIJP
Sarphatipark
AMSTERDAM-

LABEL305

Amsterdam
Kruithuisstraat 13
1018 WJ Amsterdam
+31 (0)20 261 47 49


Questions?

Now

Drinks at cafe de Polder