

A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification

Mounica Maddela and Wei Xu



THE OHIO STATE UNIVERSITY

Department of Computer Science
and Engineering

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a soft paste. It is made of apples.*

Text Simplification

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a soft paste. It is made of apples.*



Applications

- Reading assistance for children, non-native speakers and disabled.
- Improve other NLP tasks (MT, summarization ...)

Assessing **word complexity** is vital!

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a soft paste. It is made of apples.*

Assessing **word complexity** is vital!

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a soft paste. It is made of apples.*

Complex Word Identification

Assessing **word complexity** is vital!

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a **soft paste**. It is made of apples.*

liquidized sauce

thick liquid

Complex Word Identification - Substitution Generation

Assessing word complexity is vital!

INPUT: *Applesauce is a puree made of apples.*

OUTPUT: *Applesauce is a soft paste. It is made of apples.*

thick liquid

liquidized sauce

complex



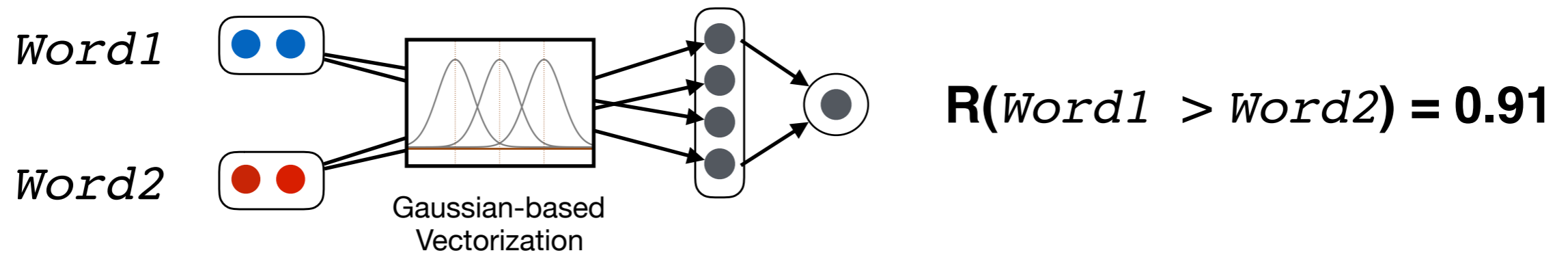
Complex Word Identification - Substitution Generation - Substitution Ranking

A Large Word-complexity Lexicon

- 15,000 English words w/ human ratings

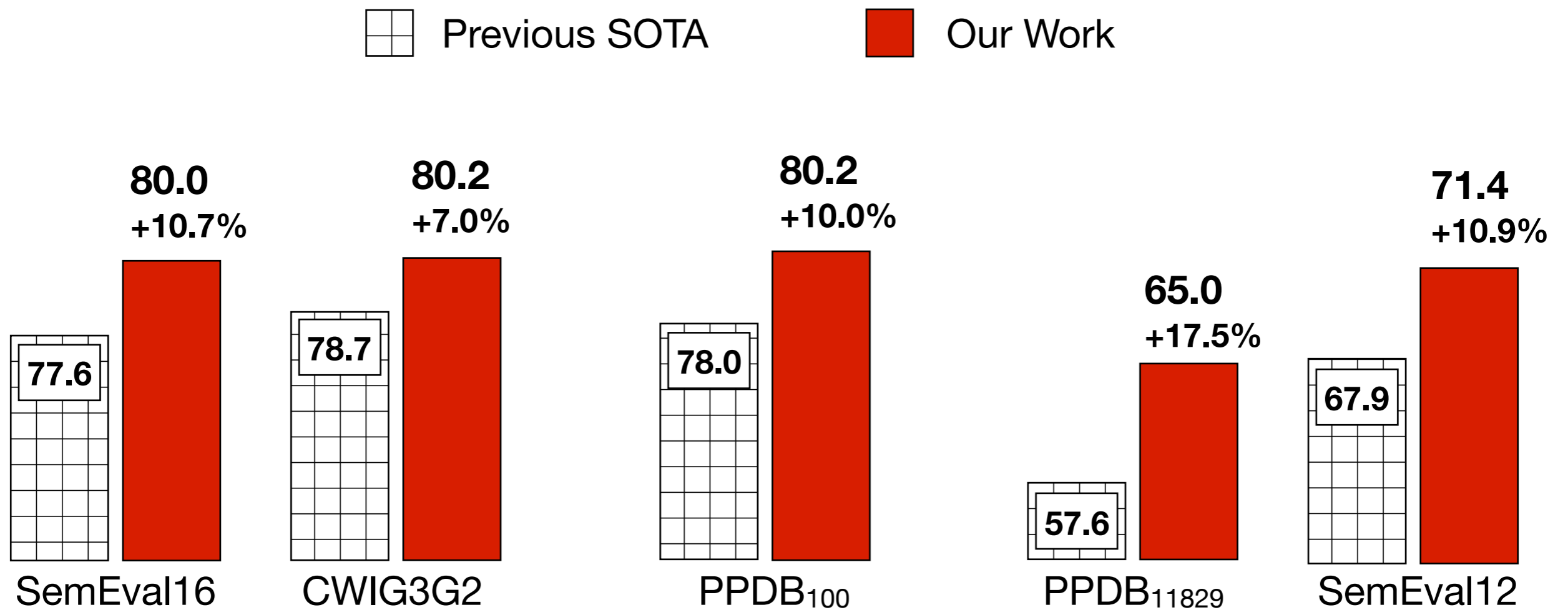
<i>day</i>	1.0		MIN 1 (simple)
<i>convenient</i>	2.4		
<i>transmitted</i>	3.2		
<i>cohort</i>	4.3		
<i>assay</i>	5.8		MAX 6 (complex)

- predict relative complexity for any given words or phrases



A Pairwise Neural Ranking Model

- improve the state-of-the-art significantly for all lexical simplification tasks



Complex Word Identification - Substitution Generation - Substitution Ranking

(% is relative error reduction)

Previous Work

Rely on **heuristics and corpus level features** to measure word complexity

- **Word length**

(Shardlow 2013, Biran et. al. 2011, and many others)

- **Word frequency in corpus**

(Bott et. al. 2011, Kajiwara et. al. 2013, Horn et. al. 2014, and many others)

- **Language model probability**

(Glavas & Stajner 2015, Paetzold & Special 2016/17, and many others)

Weakness of Previous Work

Assumption #1: shorter words are simpler

Wrong!
(21% of time*)

duly > *thoroughly*
pundit > *professional*
alien > *stranger*

* based on 2272 lexical paraphrases sampled from PPDB

Weakness of Previous Work

Assumption #2: more frequent words are simpler

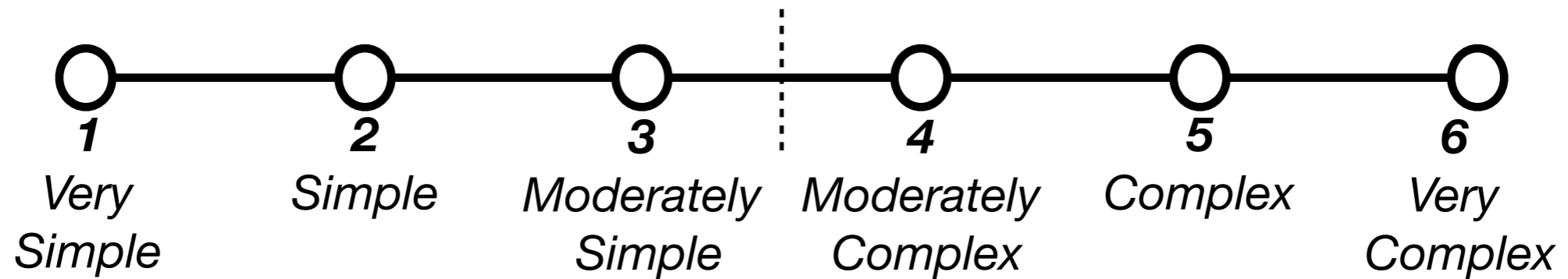
Wrong!
(14% of time*)

folly > *foolishness*
scheme > *outline*
distress > *discomfort*

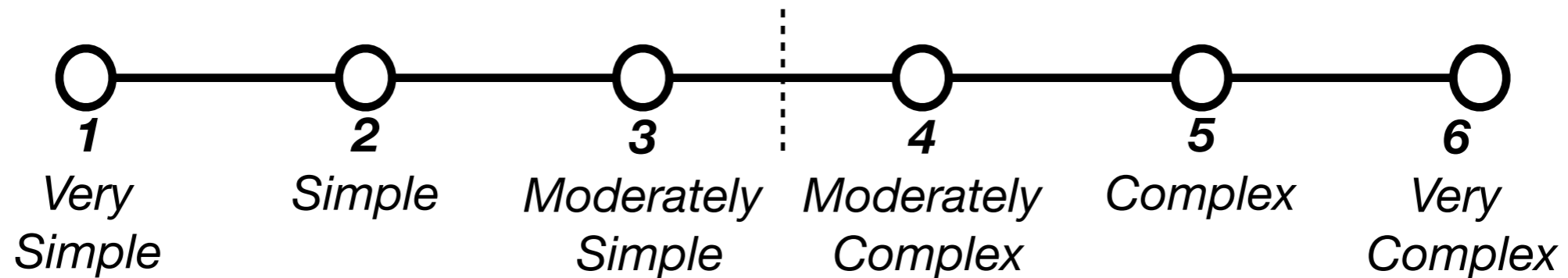
* based on 2272 lexical paraphrases sampled from PPDB

A Large Word-complexity Lexicon

- 15,000 most frequent English words from Google 1T ngram corpus
- Rated on a 6-point Likert scale



- 15,000 most frequent English words from Google 1T ngram corpus
- Rated on a 6-point Likert scale



- ▶ 11 annotators (non-native speakers)
- ▶ 5 ~ 7 ratings for each word
- ▶ 2.5 hours to rate 1000 words



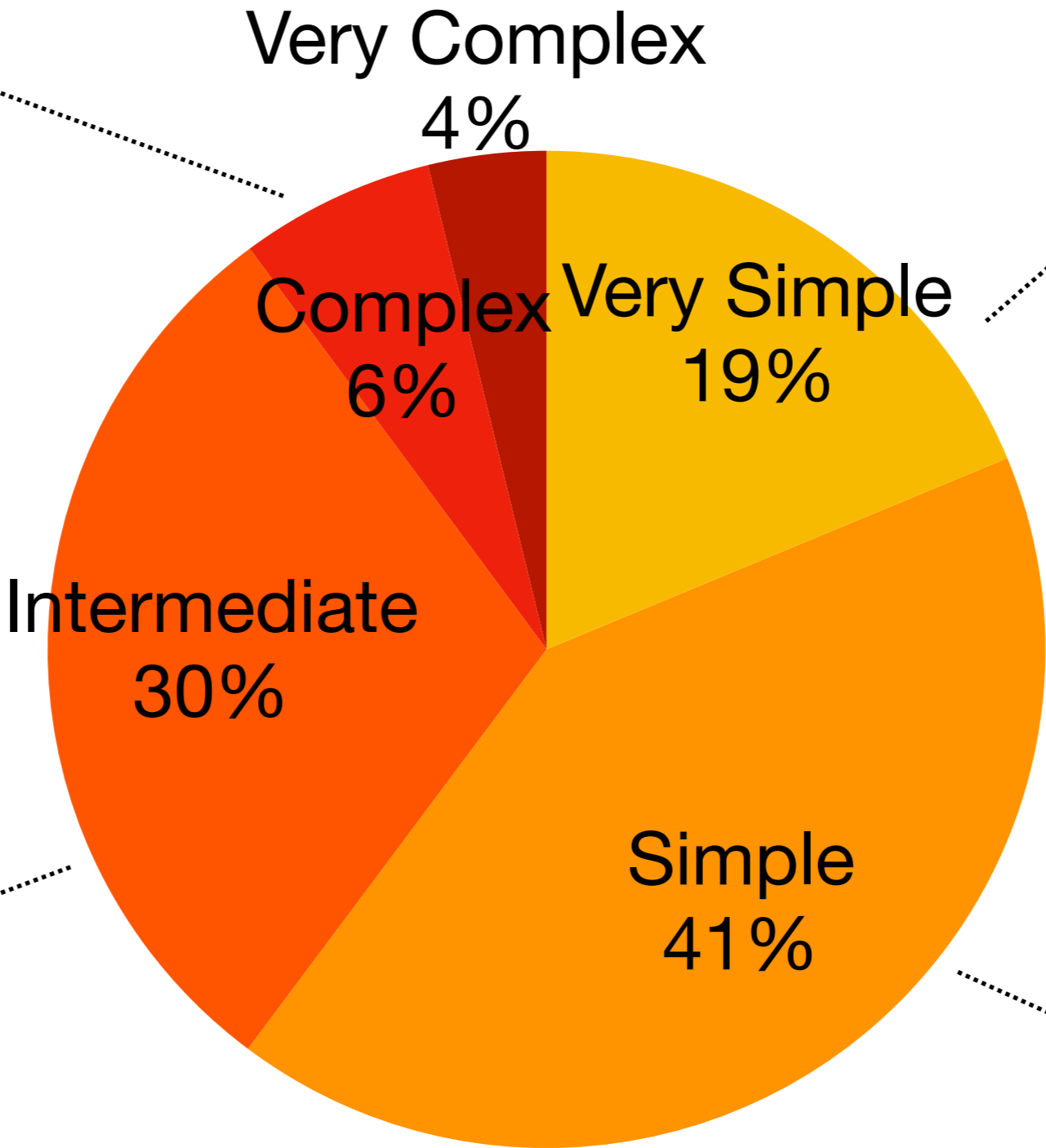
*hath
gnome
cohort
beacon
scrutiny
activism
stochastic
humanitarian
accountability*

*voyeur
swivel
claimant
facsimile
symposium*

*eat
app
dude
moon
crash
summer
yesterday*

*ion
crisis
thrust
priority
splendid
perimeter
technology
inspirational
commissioner*

*knit
cell
adjust
escape
excited
disease
pleasure
celebration
government*



- Inter-annotator agreement is 0.64 (Pearson correlation)
- One annotator rating vs. mean of the rest

Word	Score	A1	A2	A3	A4	A5
<i>muscles</i>	1.6	2	1	2	2	1
<i>pattern</i>	2.4	2	3	1	1	3
<i>educational</i>	3.2	3	3	3	3	4
<i>cortex</i>	4.2	4	4	4	4	5
<i>assay</i>	5.8	6	6	6	5	6

difference
 (one vs. rest)

- < **0.5** for **47%** of annotations
- < **1.0** for **78%** of annotations
- < **1.5** for **93%** of annotations

Evaluation* - Complex Word Identification

- Complex Word Identification Shared Task - BEA@NAACL'18
- 34879 sentences from Wikipedia and news articles
- 27299 training, 3328 development, 4252 test instances

Input	<i>The whale was sensing him with sound pulses.</i>
Output	<i>[Complex, simple]</i>

* see paper for full evaluation on 3 lexical simplification tasks and 5 benchmark datasets

Evaluation

- Complex Word Identification Shared Task 2018
- 27299 training, 3328 development, 4252 test instances

	F-score	Accuracy
Senses	62.3	54.1
SimpleWiki Frequency	63.3	61.6
Length	65.9	67.7
(Yimam et al. 2017)	66.6	76.7
(Paetzold et al. 2016)	73.8	78.7

Evaluation

- Complex Word Identification Shared Task 2018
- 27299 training, 3328 development, 4252 test instances

	F-score	Accuracy
Senses	62.3	54.1
SimpleWiki Frequency	63.3	61.6
Length	65.9	67.7
(Yimam et al. 2017)	66.6	76.7
(Paetzold et al. 2016)	73.8	78.7
Our Lexicon	67.5	69.8

Annotations: A red arrow points from 65.9 to 67.5 with a label +1.6. Another red arrow points from 67.7 to 69.8 with a label +2.1.

Evaluation

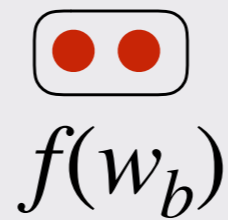
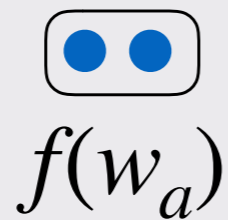
- Complex Word Identification Shared Task 2018
- 27299 training, 3328 development, 4252 test instances

	F-score	Accuracy
Senses	62.3	54.1
SimpleWiki Frequency	63.3	61.6
Length	65.9	67.7
(Yimam et al. 2017)	66.6	76.7
(Paetzold et al. 2016)	73.8	78.7
Our Lexicon	67.5	69.8
(Yimam et al. 2017) + Our Lexicon	*68.8	*78.1
(Paetzold et al. 2016) + Our Lexicon	*74.8	*80.2

* statistically significant ($p < 0.01$) based on the paired bootstrap test

A Pairwise Neural Ranking Model

Feature
Extraction



Input Word/Phrase Pair

$\langle w_a : \text{adversary} , w_b : \text{enemy} \rangle$

Word-Complexity Lexicon Score

0/1 binary indicator

word length
word frequency
number of syllables
ngram probabilities

Feature
Extraction

$f(w_a)$

$f(w_b)$

Input Word/Phrase Pair

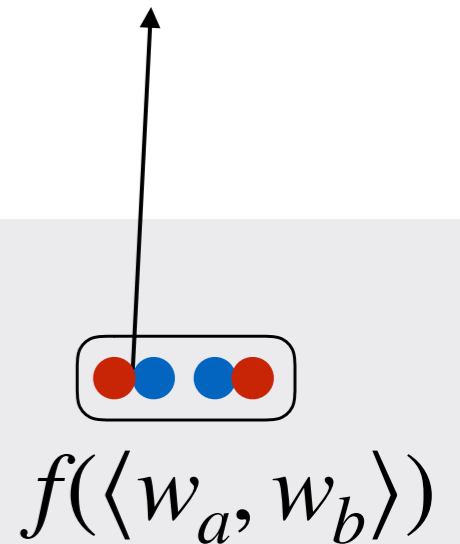
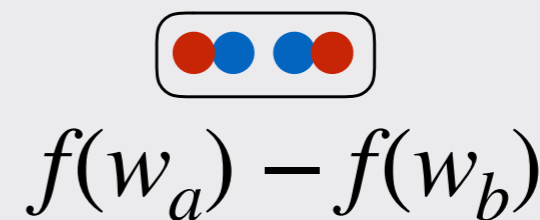
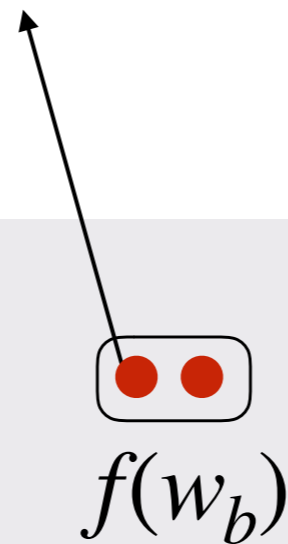
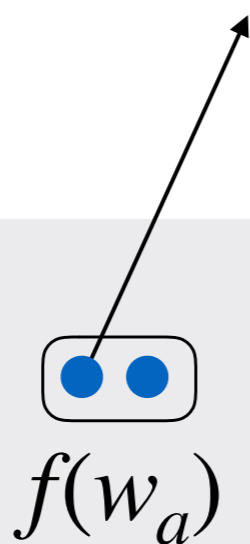
$\langle w_a : \text{adversary} , w_b : \text{enemy} \rangle$

Word-Complexity Lexicon Score 0/1 binary indicator

word length
word frequency
number of syllables
ngram probabilities

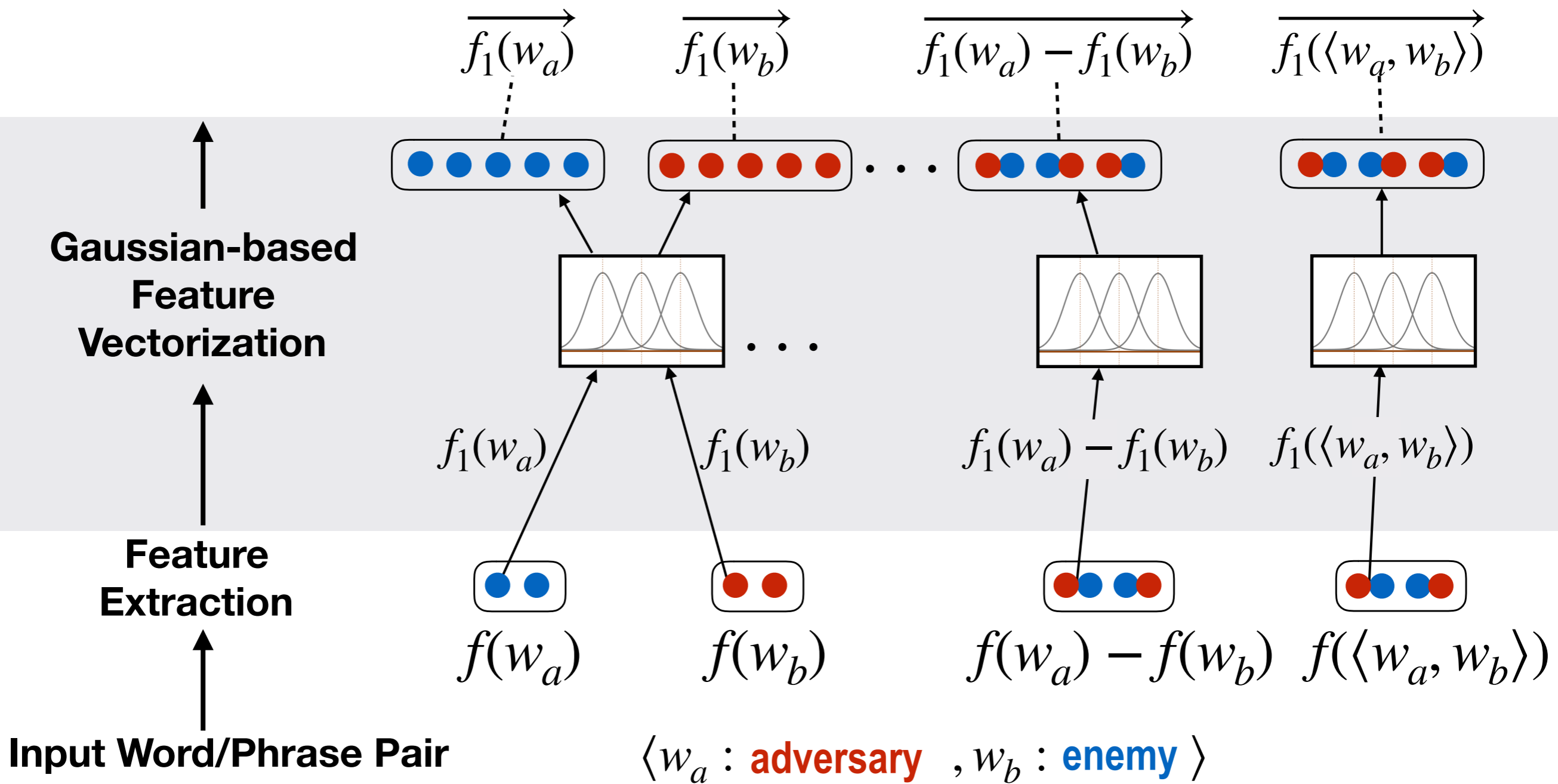
PPDB paraphrase score
word2vec cosine similarity

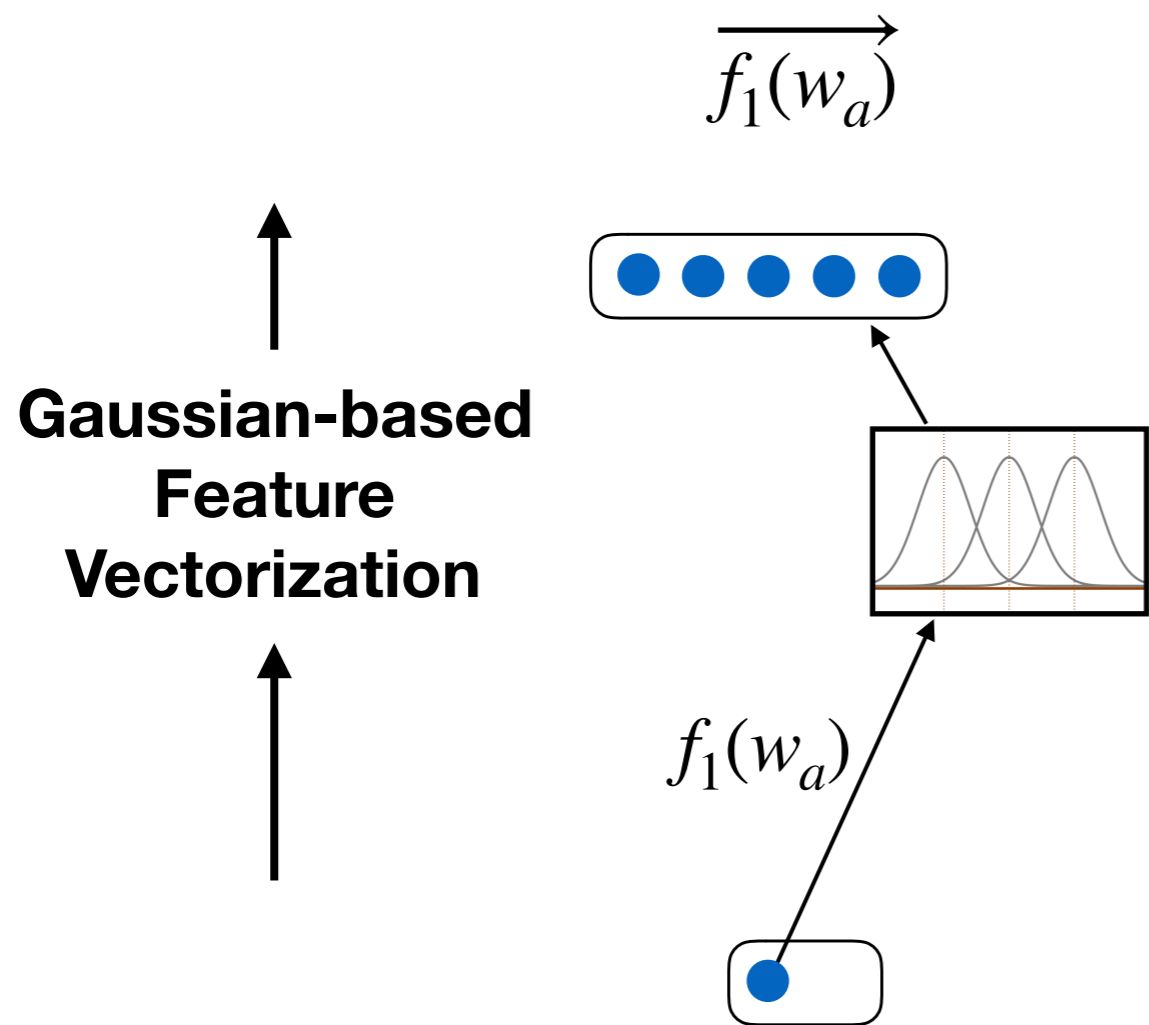
Feature
Extraction



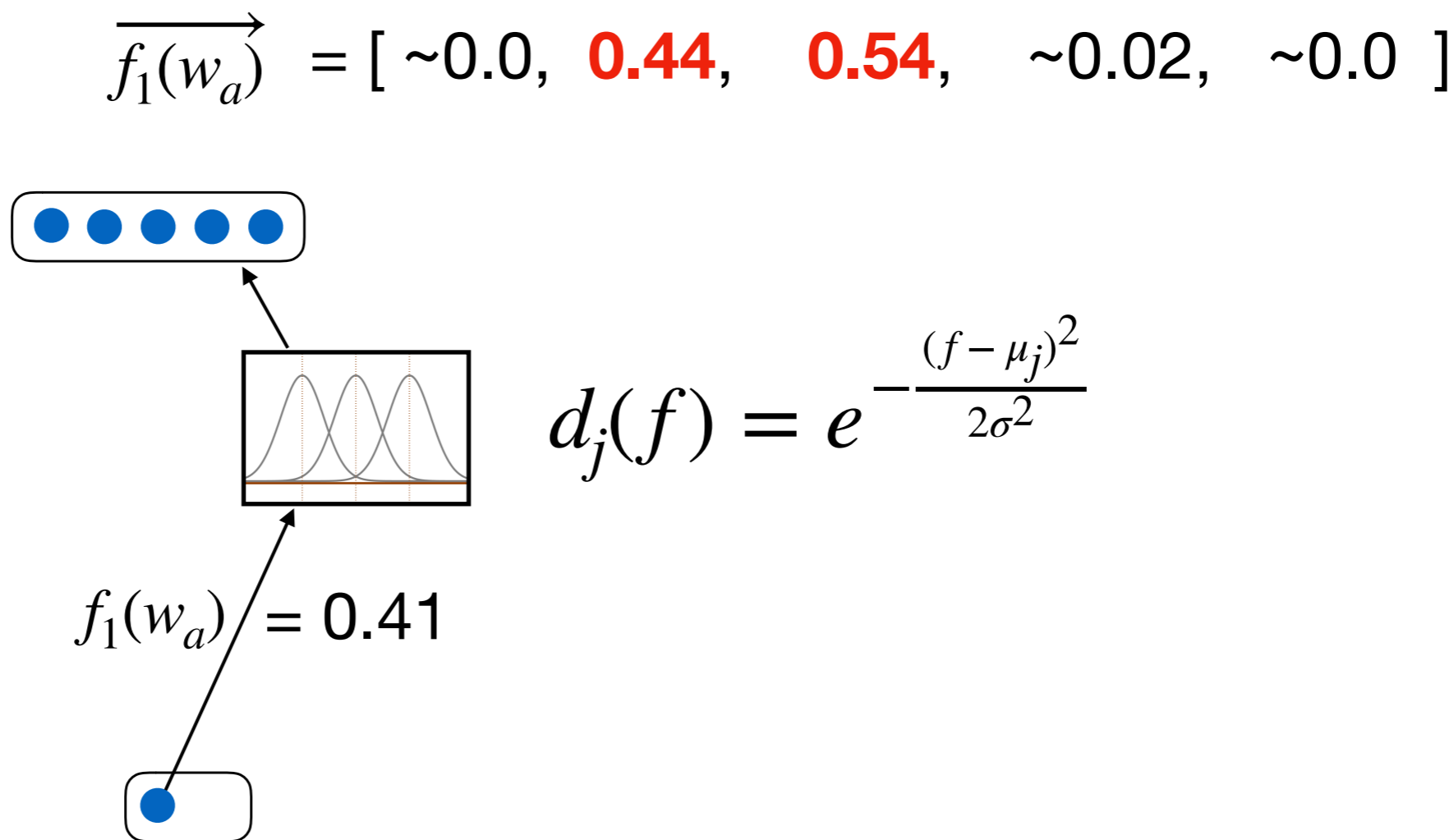
Input Word/Phrase Pair

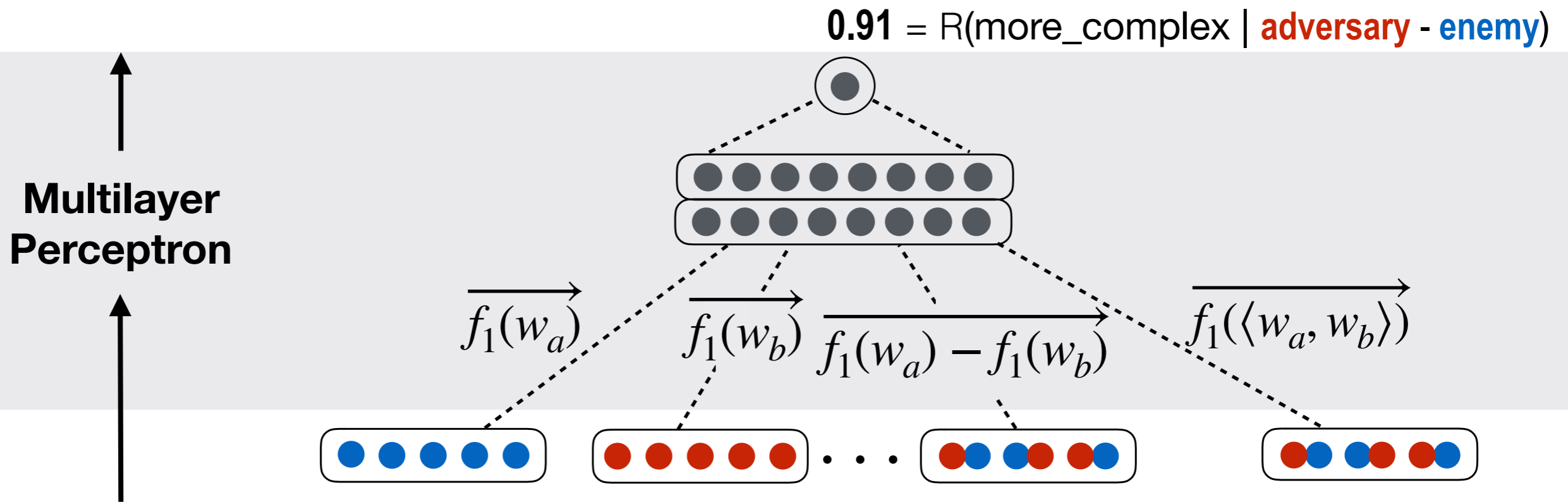
$\langle w_a : \text{adversary} , w_b : \text{enemy} \rangle$





Gaussian-based
Feature
Vectorization





$R > 0 \Rightarrow w_a$ is more complex than w_b

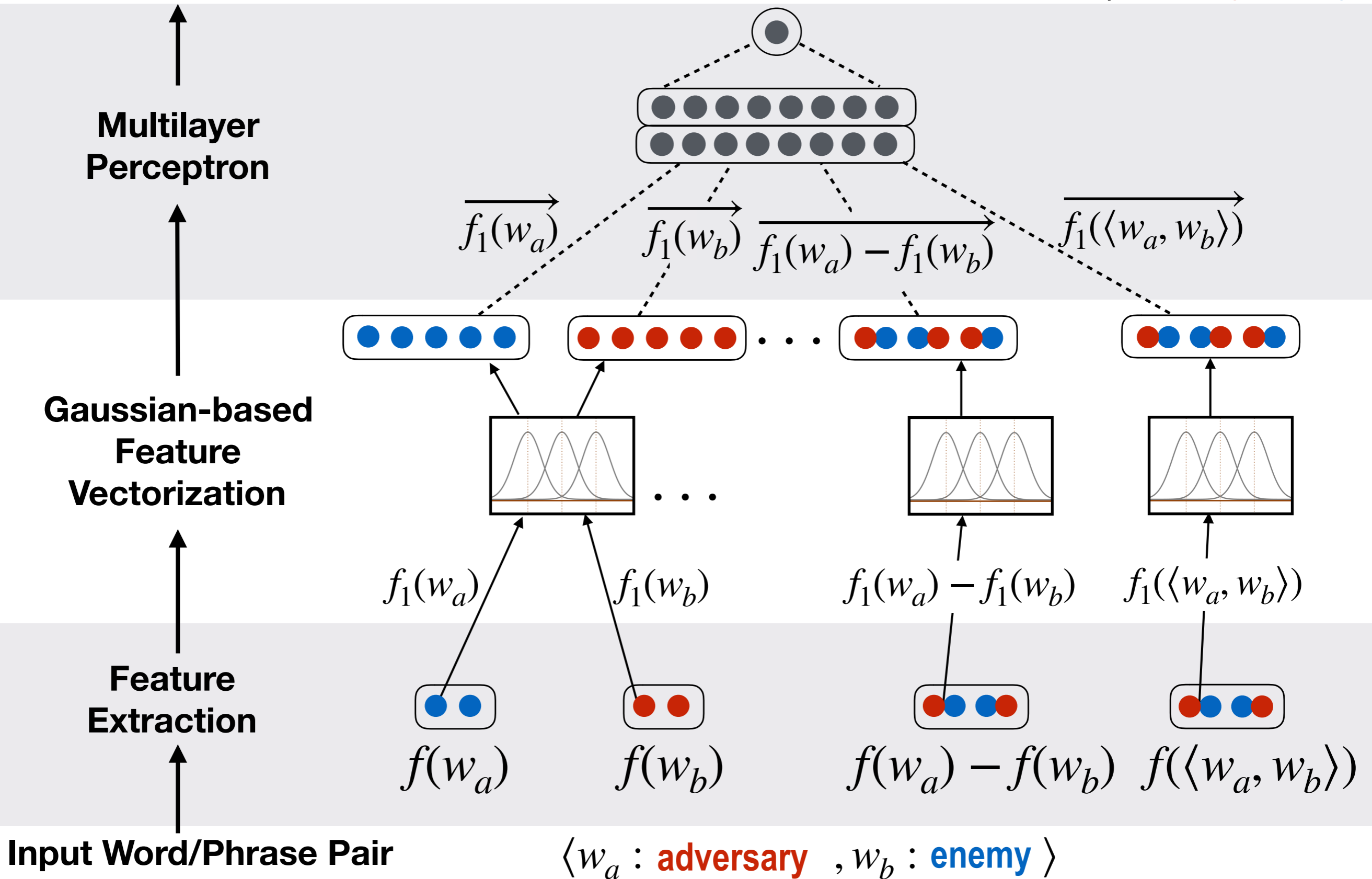
$R < 0 \Rightarrow w_a$ is simpler than w_b

$|R|$ indicates complexity difference

$\langle w_a : \text{adversary} , w_b : \text{enemy} \rangle$

Neural Readability Ranking Model

$$0.91 = R(\text{more_complex} \mid \text{adversary} - \text{enemy})$$



Evaluation**

- English Lexical Simplification Shared Task - SemEval 2012
- 300 training sentences, 1710 test sentences

Input	<i>There were also pieces that would have been <u>terrible</u> in any environment.</i>
(Paetzold & Specia 2017)	<i>awful, very bad, dreadful</i>
Our Model + Our Lexicon	<i>very bad, awful, dreadful</i>
Gold truth	<i>very bad, awful, dreadful</i>

** see paper for full evaluation on 3 lexical simplification tasks and 5 benchmark datasets

Evaluation

- English Lexical Simplification Shared Task - SemEval 2012
- 300 training sentences, 1710 test sentences

		Precision@1	Pearson
heuristics	(Biran et al. 2011)	51.3	0.505
SVM	(Jauhar & Specia 2012)	60.2	0.575
heuristics	(Kajiwara et al. 2013)	60.4	0.649
SVM	(Horn et al. 2014)	63.9	0.673
heuristics	(Glavaš & Štajner 2015)	63.2	0.644
SVM	(Paetzold & Specia 2015)	65.3	0.677
neural	(Paetzold & Specia 2017)	65.6	0.679
neural	Our Model + Lexicon + Gaussian	67.3*	0.714*

+0.2 (from 65.3 to 65.6)
+0.002 (from 0.677 to 0.679)
+1.7 (from 65.6 to 67.3)
+0.035 (from 0.679 to 0.714)

* statistically significant ($p < 0.05$) based on the paired bootstrap test

Evaluation

- English Lexical Simplification Shared Task - SemEval 2012
- 300 training sentences, 1710 test sentences

		Precision@1	Pearson
heuristics	(Biran et al. 2011)	51.3	0.505
SVM	(Jauhar & Specia 2012)	60.2	0.575
heuristics	(Kajiwara et al. 2013)	60.4	0.649
SVM	(Horn et al. 2014)	63.9	0.673
heuristics	(Glavaš & Štajner 2015)	63.2	0.644
SVM	(Paetzold & Specia 2015)	65.3	0.677
neural	(Paetzold & Specia 2017)	65.6	0.679
neural	Our Model + Gaussian	66.6	0.702*
neural	Our Model + Lexicon + Gaussian	67.3*	0.714*

+0.2 (from 65.3 to 65.6)
 +0.002 (from 0.677 to 0.679)
 +1.7 (from 65.6 to 66.6)
 +0.035 (from 0.679 to 0.702)

* statistically significant ($p < 0.05$) based on the paired bootstrap test

Evaluation

- English Lexical Simplification Shared Task - SemEval 2012
- 300 training sentences, 1710 test sentences

		Precision@1	Pearson
heuristics	(Biran et al. 2011)	51.3	0.505
SVM	(Jauhar & Specia 2012)	60.2	0.575
heuristics	(Kajiwara et al. 2013)	60.4	0.649
SVM	(Horn et al. 2014)	63.9	0.673
heuristics	(Glavaš & Štajner 2015)	63.2	0.644
SVM	(Paetzold & Specia 2015)	65.3	0.677
neural	(Paetzold & Specia 2017)	65.6	0.679
neural	Our Model	65.4	0.682
neural	Our Model + Gaussian	66.6	0.702*
neural	Our Model + Lexicon + Gaussian	67.3*	0.714*

+0.2 (from 65.3 to 65.6)
 +0.002 (from 0.677 to 0.679)
 +1.7 (from 65.6 to 67.3)
 +0.035 (from 0.679 to 0.714)

* statistically significant ($p < 0.05$) based on the paired bootstrap test

Evaluation - Error Analysis

Input	<i>The colonies of one <u>strain</u> appeared smooth.</i>
(Paetzold & Specia 2017)	<i>sort, type, breed, variety</i>
Our Model + Our Lexicon	<i>type, sort, breed, variety</i>
Gold truth	<i>type, sort, variety, breed</i>

Input	<i>No damage or <u>casualties</u> were reported.</i>
(Paetzold & Specia 2017)	<i>injuries, accidents, deaths, fatalities</i>
Our Model + Our Lexicon	<i>injuries, deaths, accidents, fatalities</i>
Gold truth	<i>deaths, injuries, accidents, fatalities</i>


SimplePPDB++

- 14.1 million paraphrase rules w/ improved complexity ranking scores

Paraphrase Rule		Score
	→ <i>self-supporting</i>	0.93
<i>self-reliant</i>	→ <i>self-sufficient</i>	0.48
	→ <i>self-sustainable</i> complex	-0.60
	→ <i>possible</i>	0.94
<i>viable</i>	→ <i>realistic</i>	0.15
	→ <i>plausible</i>	-0.91
	→ <i>in-depth review</i>	0.89
<i>detailed assesement</i>	→ <i>careful examination</i>	0.28
	→ <i>comprehensive evaluation</i>	-0.87

Thanks

- **Word-Complexity Lexicon & SimplePPDB++** are available!

<i>day</i>	1.0	
<i>convenient</i>	2.4	
<i>transmitted</i>	3.2	
<i>cohort</i>	4.3	
<i>assay</i>	5.8	

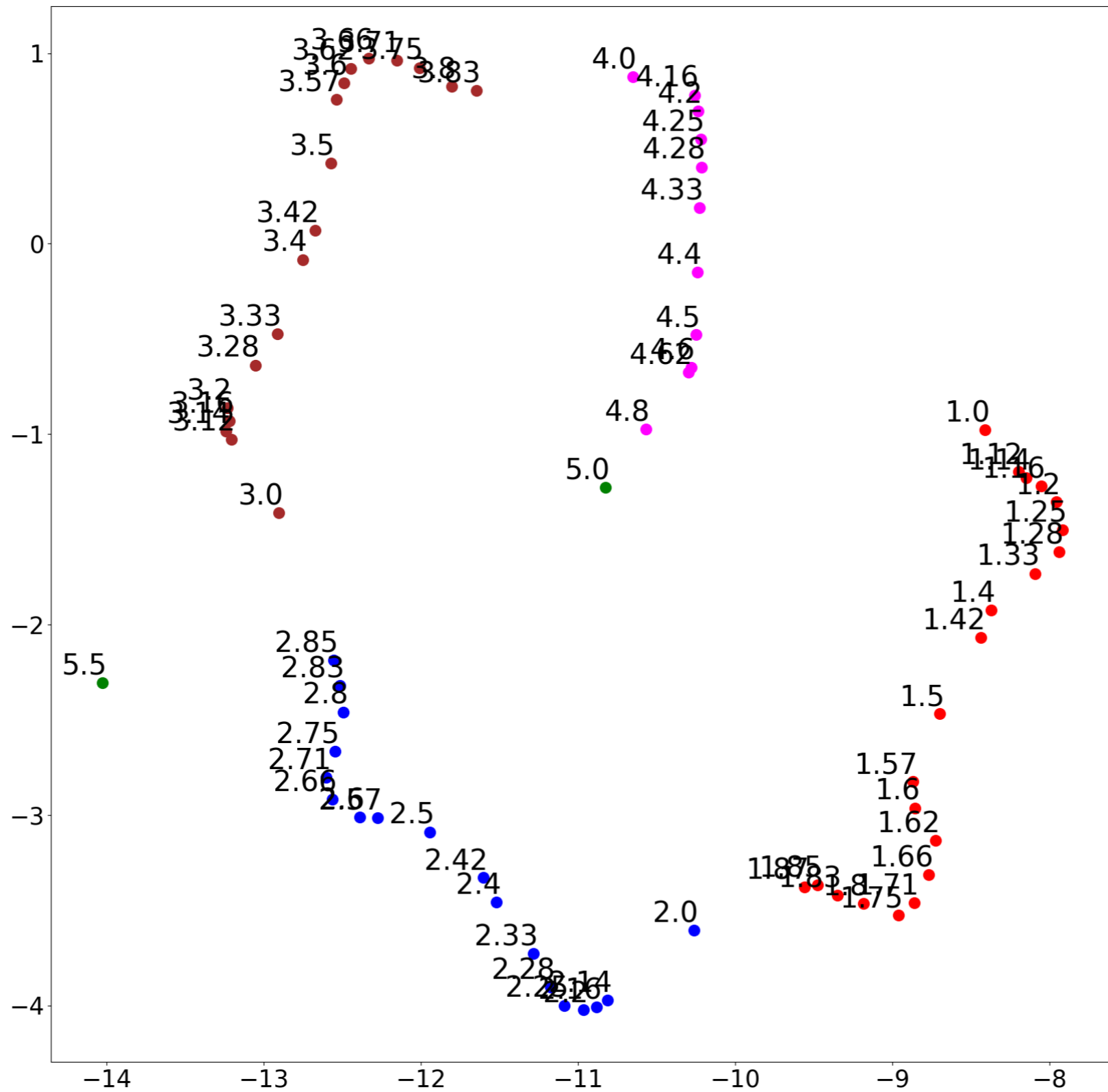
MIN 1 (simple)

MAX 6 (complex)

- PyTorch Code for the **Neural Ranking model** is also available!

https://github.com/mounicam/lexical_simplification

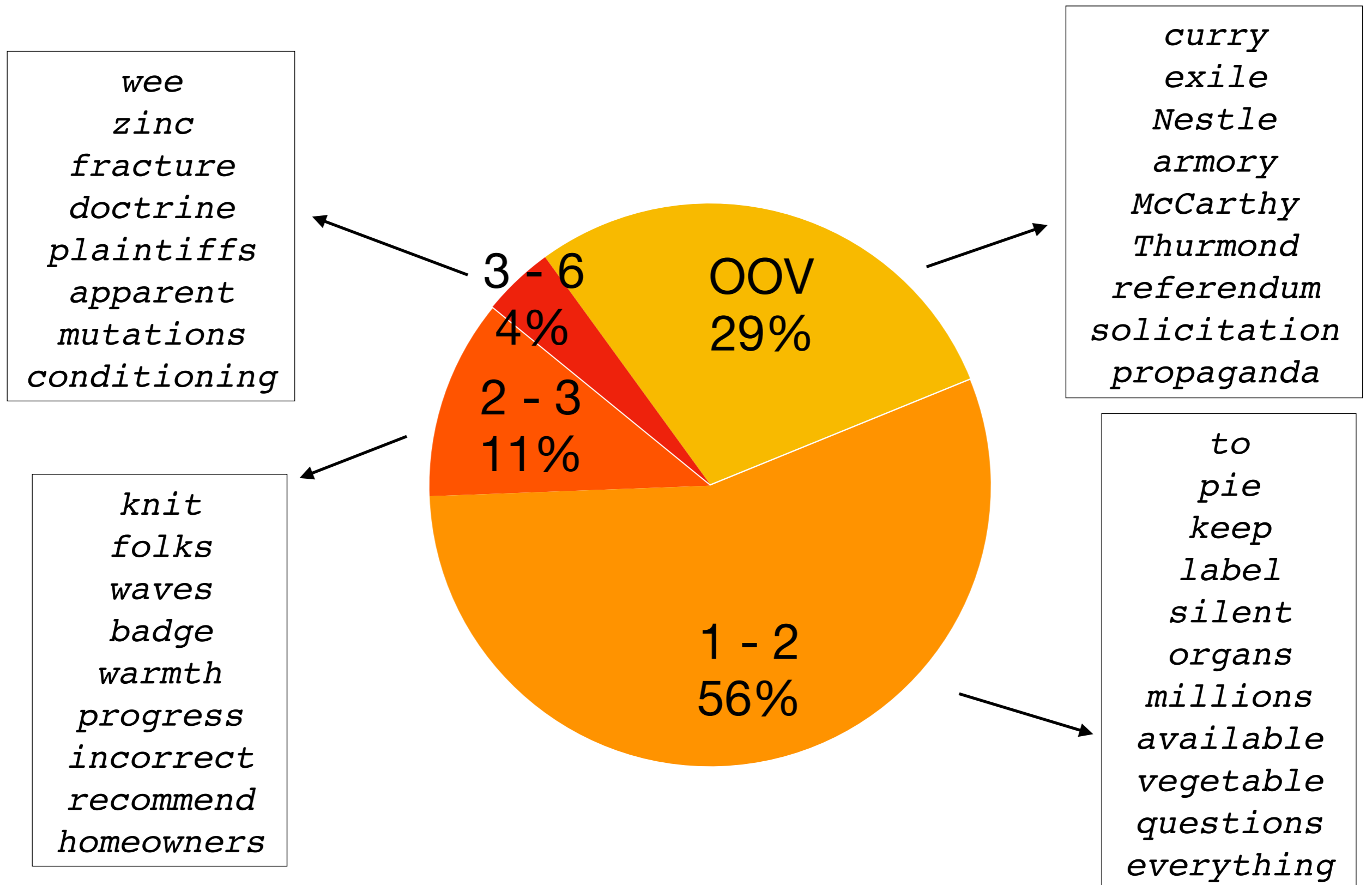
- Contacts: Mounica Maddela & Wei Xu (Ohio State University)



t-SNE visualization of the complexity scores, ranging between 1.0 and 6.0

Word-Complexity Lexicon

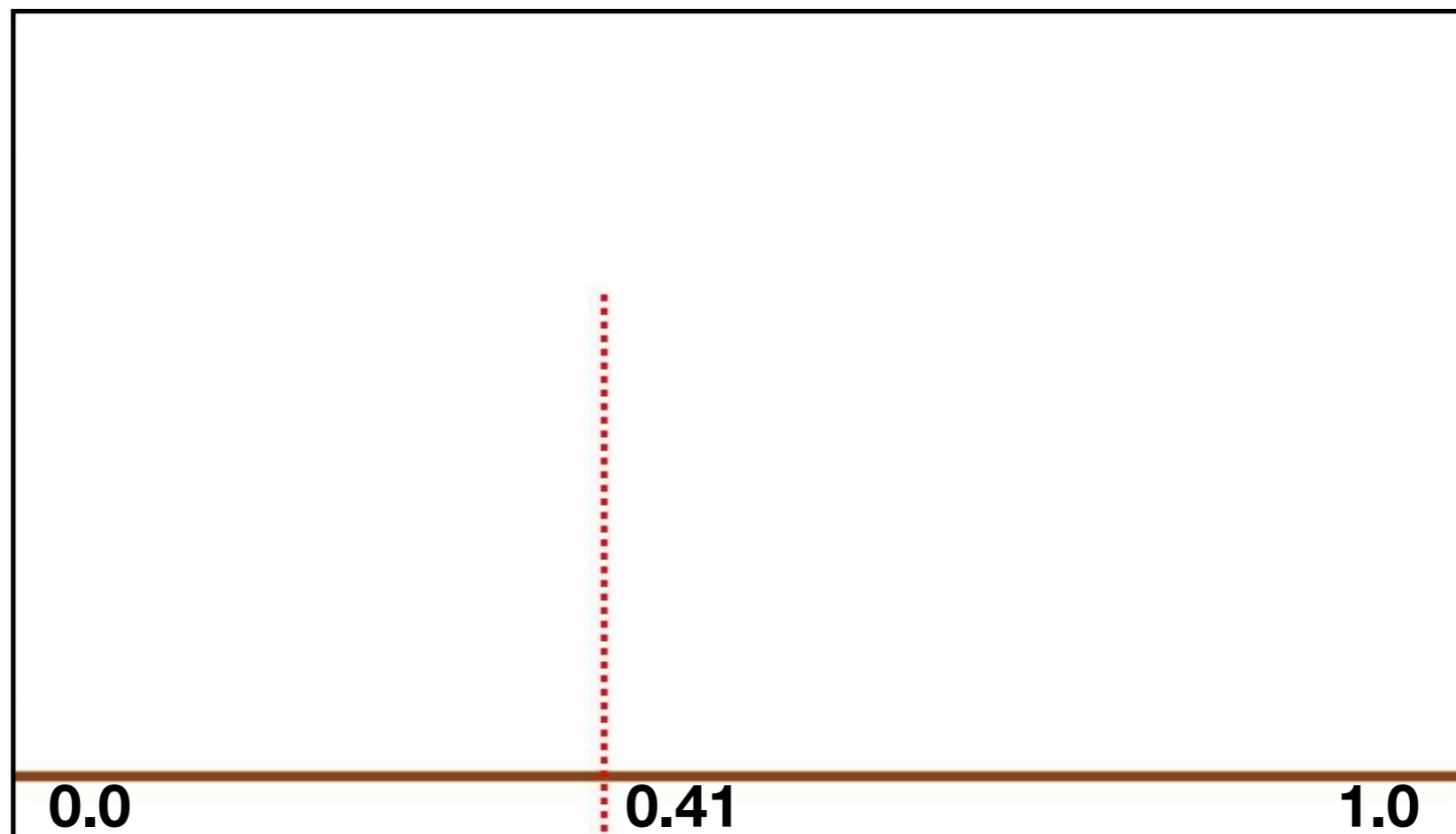
Coverage over Penn Treebank (~1.1 million words)



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

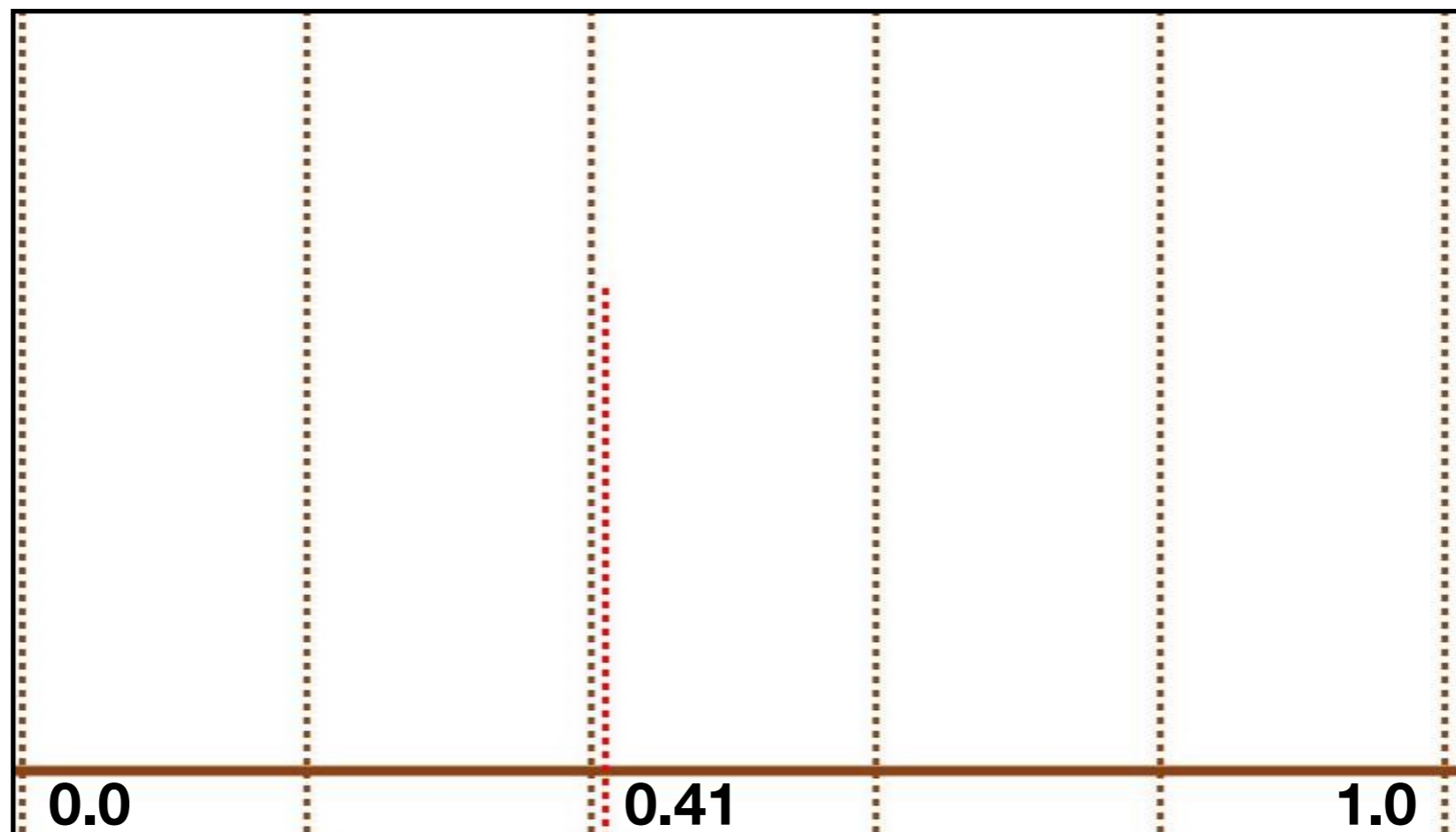
Vectorized feature : $f(w) = [\sim 0.0, 0.44, 0.54, \sim 0.02, \sim 0.0]$



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

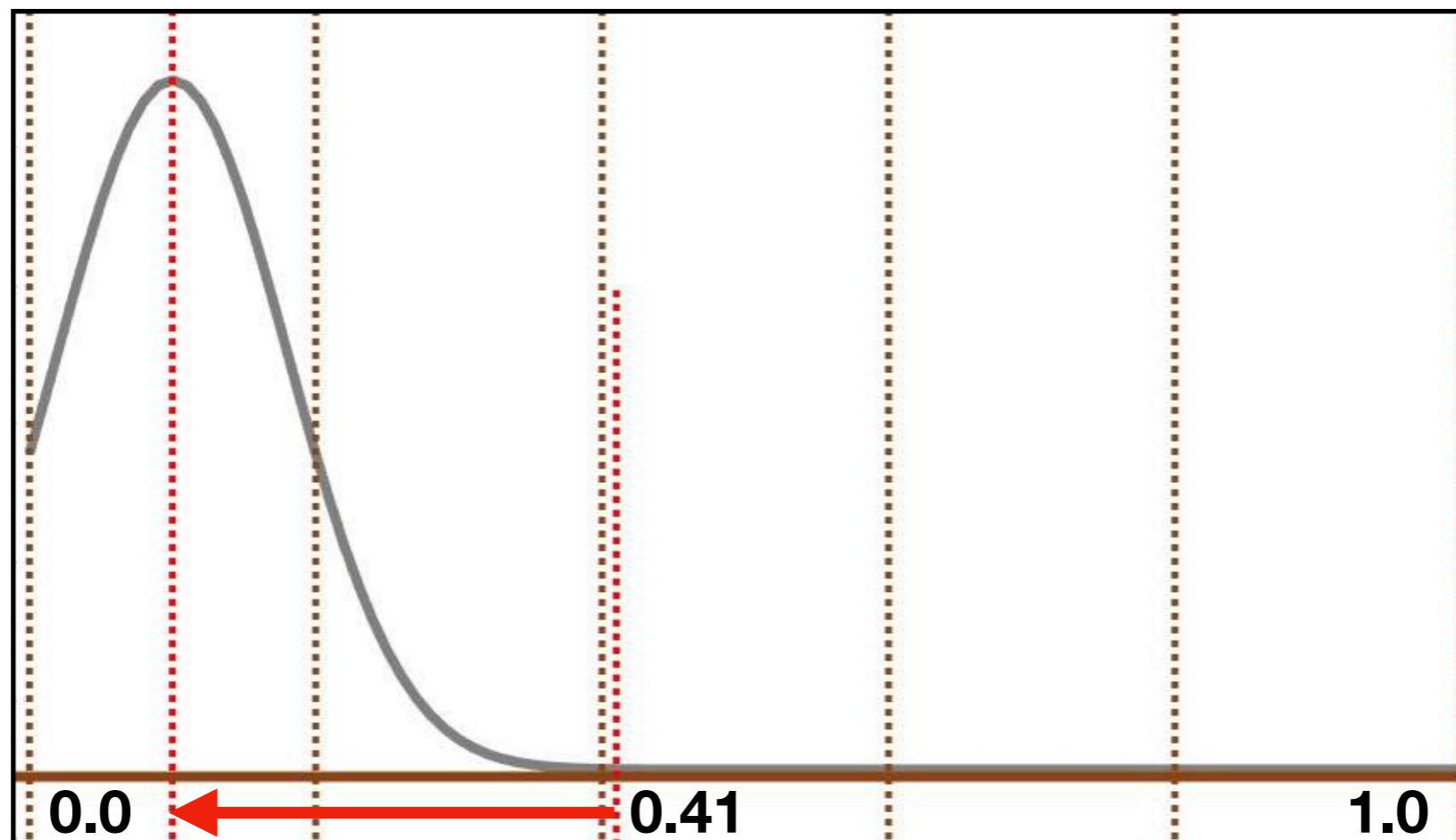
Vectorized feature : $f(w) = [\sim 0.0, 0.44, 0.54, \sim 0.02, \sim 0.0]$



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

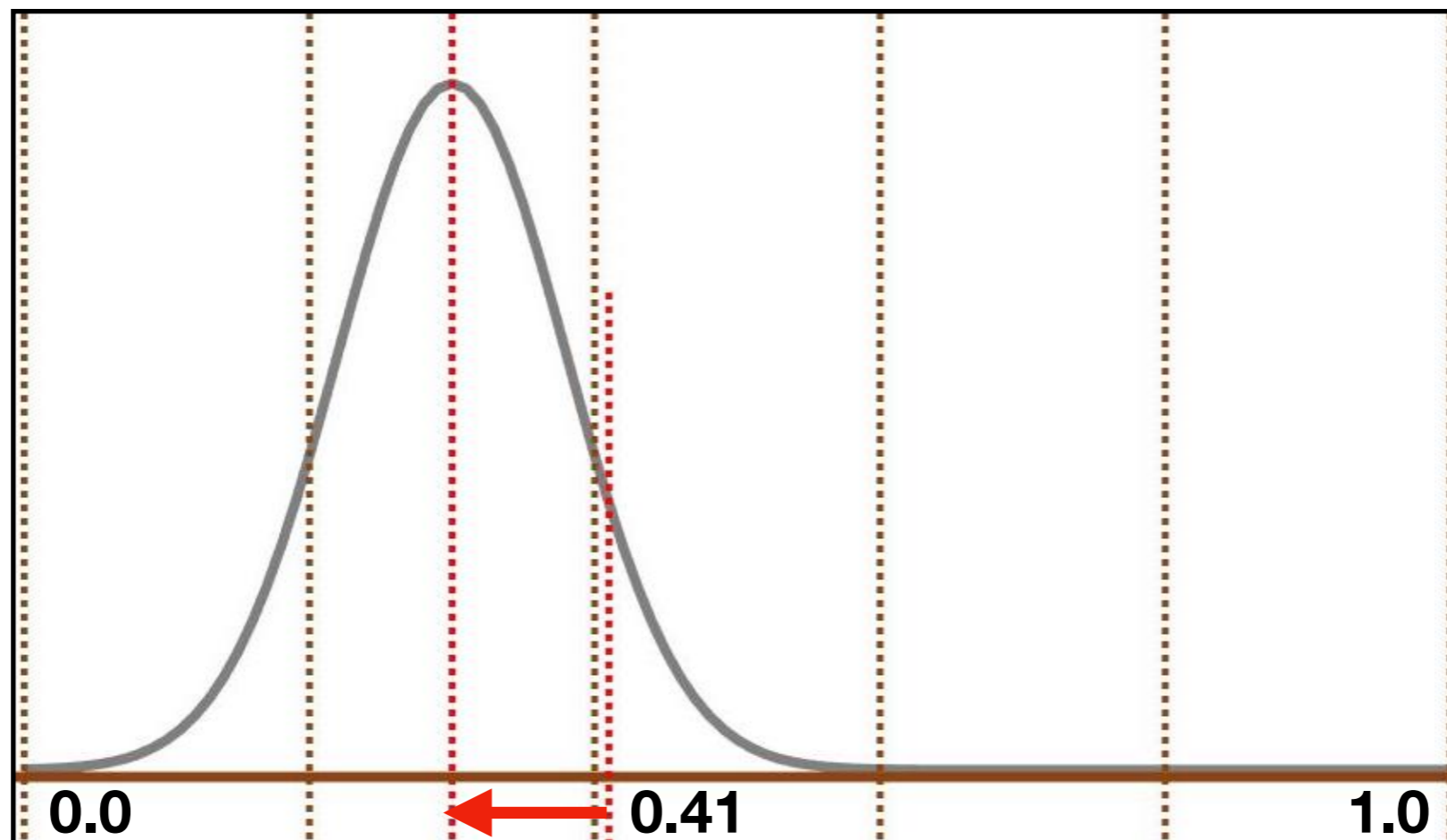
Vectorized feature : $f(w) = [\sim 0.0, \quad]$



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

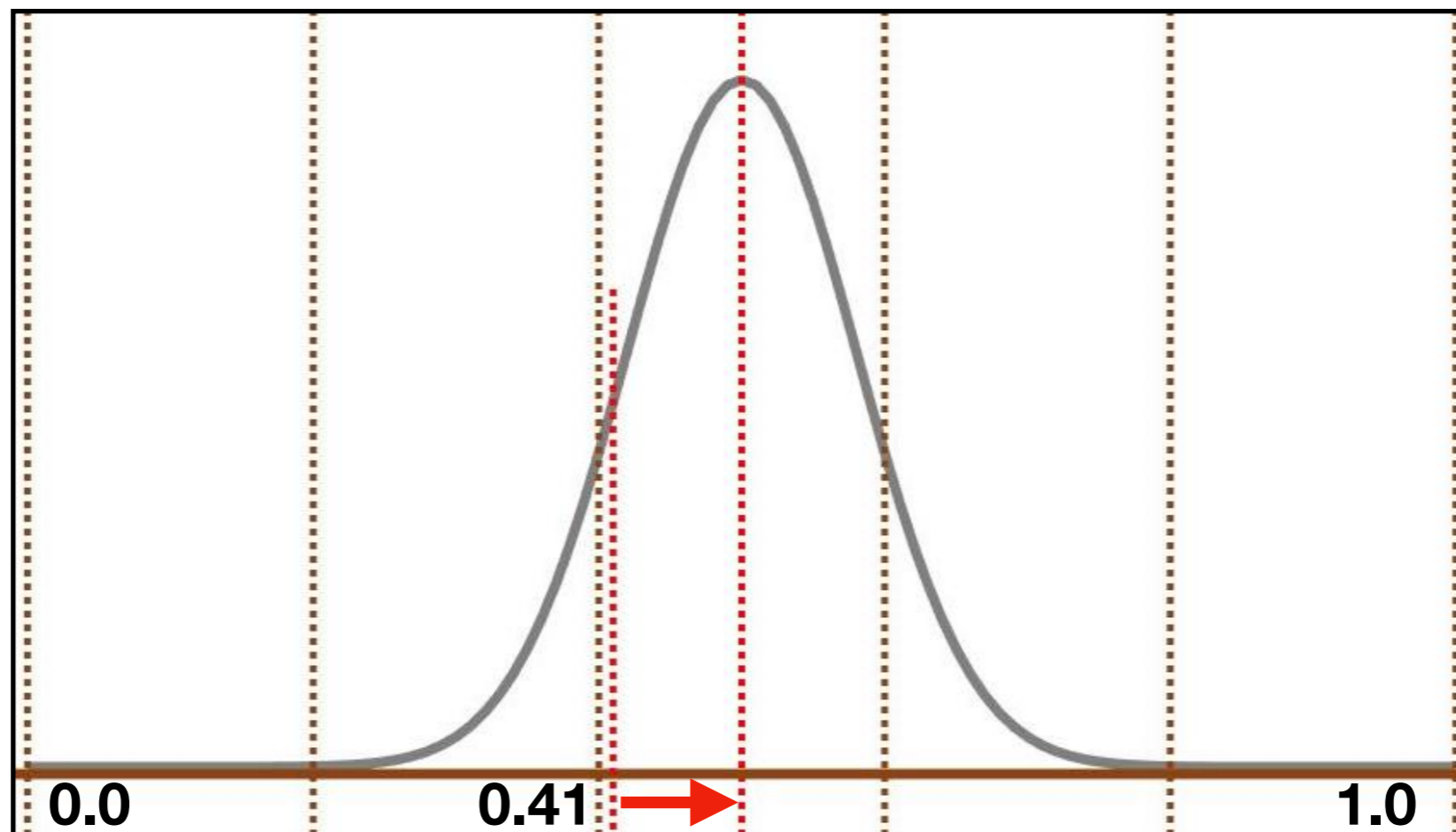
Vectorized feature : $f(w) = [\sim 0.0, \mathbf{0.44}, \dots]$



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

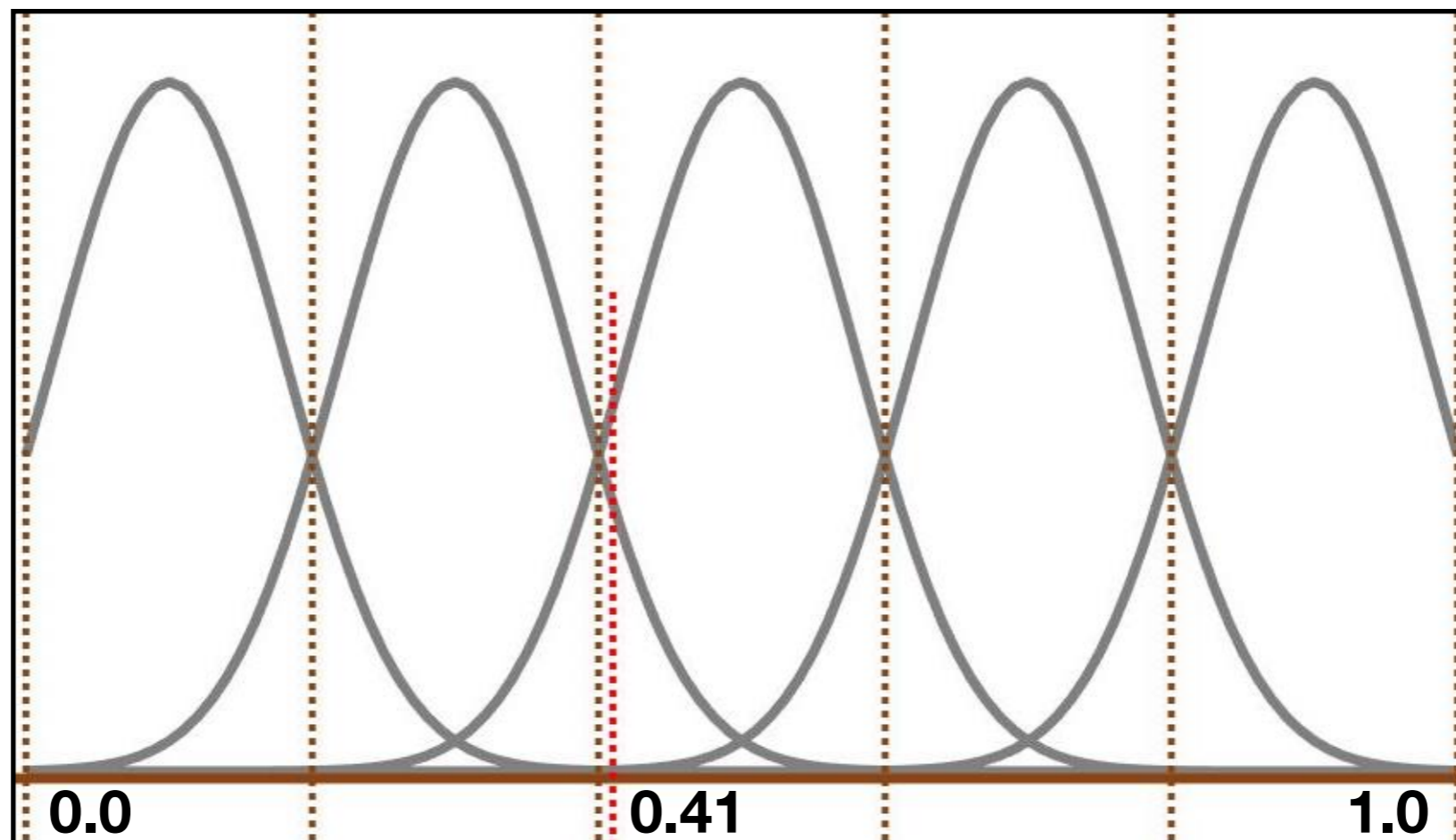
Vectorized feature : $f(w) = [\sim 0.0, 0.44, \mathbf{0.54},]$



Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

Vectorized feature : $f(w) = [\sim 0.0, 0.44, 0.54, \sim 0.02, \sim 0.0]$



Substitution Ranking - Correct Examples

- ▶ Our Model predicts the correct output

Input	<i>The <u>concept</u> of a “picture element” dates to the earliest days of television.</i>
(Paetzold & Specia 2017)	<i>theory, thought, idea</i>
Our Model + Our Lexicon	<i>idea, thought, theory</i>
Gold truth	<i>idea, thought, theory</i>

- ▶ Our Model handles phrases better than previous SOTA.

Input	<i>There were also pieces that would have been <u>terrible</u> in any environment.</i>
(Paetzold & Specia 2017)	<i>awful, very bad, dreadful</i>
Our Model + Our Lexicon	<i>very bad, awful, dreadful</i>
Gold truth	<i>very bad, awful, dreadful</i>