Learning for Unlimited Human Language

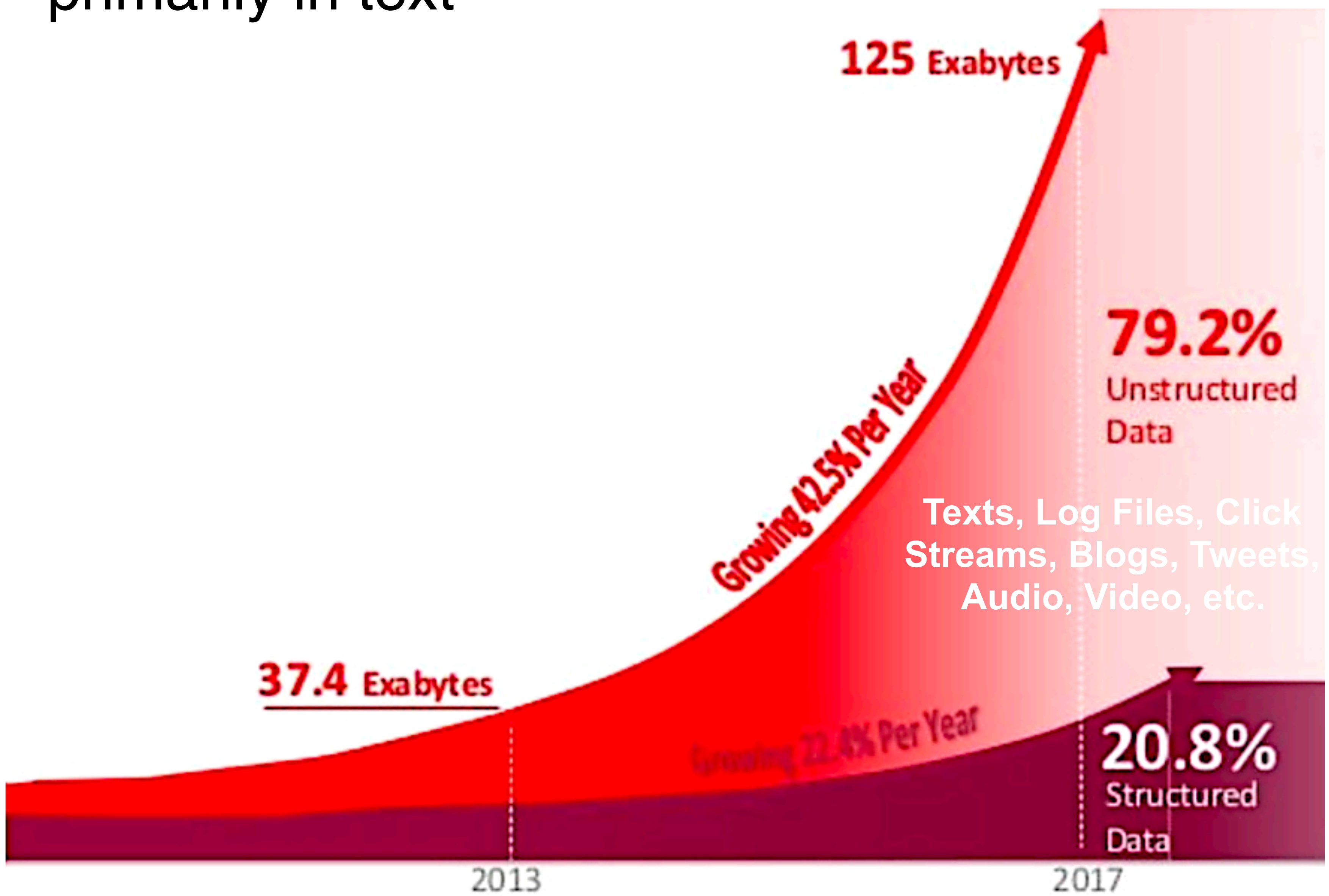
Wei Xu

THE OHIO STATE UNIVERSITY
Department of Computer Science and Engineering

with Wuwei Lan, Mounica Maddala, Alan Ritter, Chris Callison-Burch, Jeniya Tabassum, Daniel Preoţiuc-Pietro, Chaitanya Kulkarni, Hua He, Bill Dolan, Siyu Qiu, Quanze Chen

Unlimited Text in practice

80% of world information is unstructured
primarily in text



Source: ISD - 2014, Structured Data vs. Unstructured Data: The Balance of Power Continues to Shift

Unlimited Text in theory

"Almost any single (relatively complex) meaning can be implemented by an astonishingly high number of synonymous surface expressions."

Meaning-Text Linguistic Theory (Žolkovskij & Mel'čuk, 1965; ~ now)

meaning	=	invariant of paraphrases
text	=	'virtual paraphrasing'
paraphrases	=	synonymous linguistic expressions

Unlimited Text in theory

"Almost any single (relatively complex) meaning can be implemented by an astonishingly high number of synonymous surface expressions."

Meaning-Text Linguistic Theory (Žolkovskij & Mel'čuk, 1965; ~ now)

meaning	=	invariant of **paraphrases**
text	=	'virtual **paraphrasing**'
paraphrases	=	synonymous linguistic expressions

my take on Unlimited Text

learn and model very-large-scale paraphrases

wealthy	**word**	*rich*
the king's speech	**phrase**	*His Majesty's address*
… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …	**sentence**	*… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …*

What's good about Paraphrases ?

fundamentally useful for a wide range of applications

e.g. Question Answering

Who is the CEO stepping down from Boeing?

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …

What's good about Paraphrases ?

fundamentally useful for a wide range of applications

e.g. Question Answering

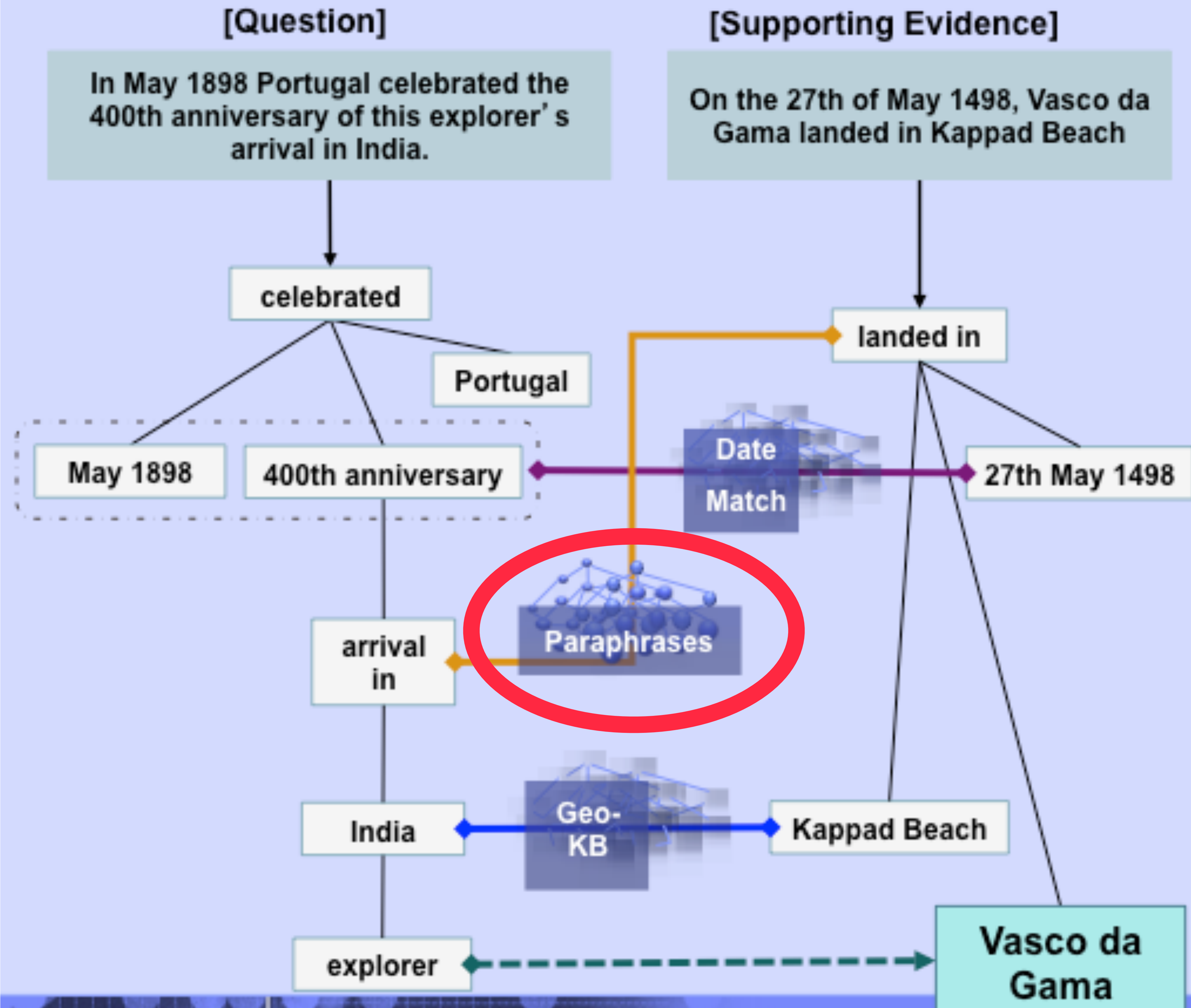
Who is the CEO stepping down from Boeing?

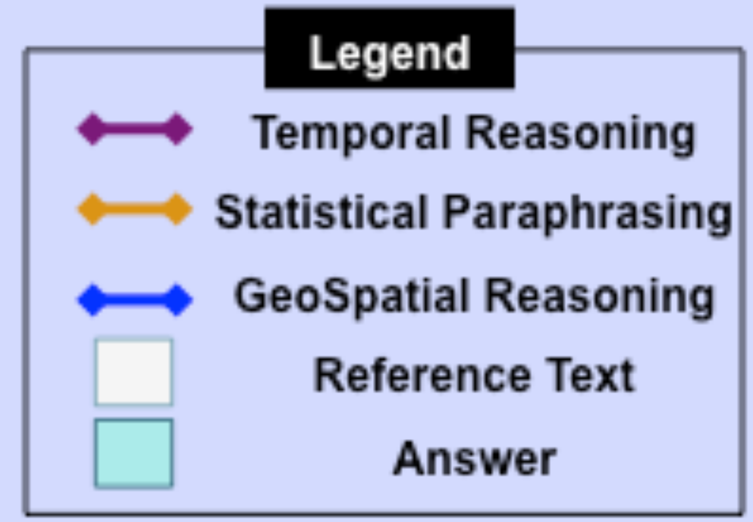
match

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …

Watson leverages multiple algorithms to perform deeper analysis



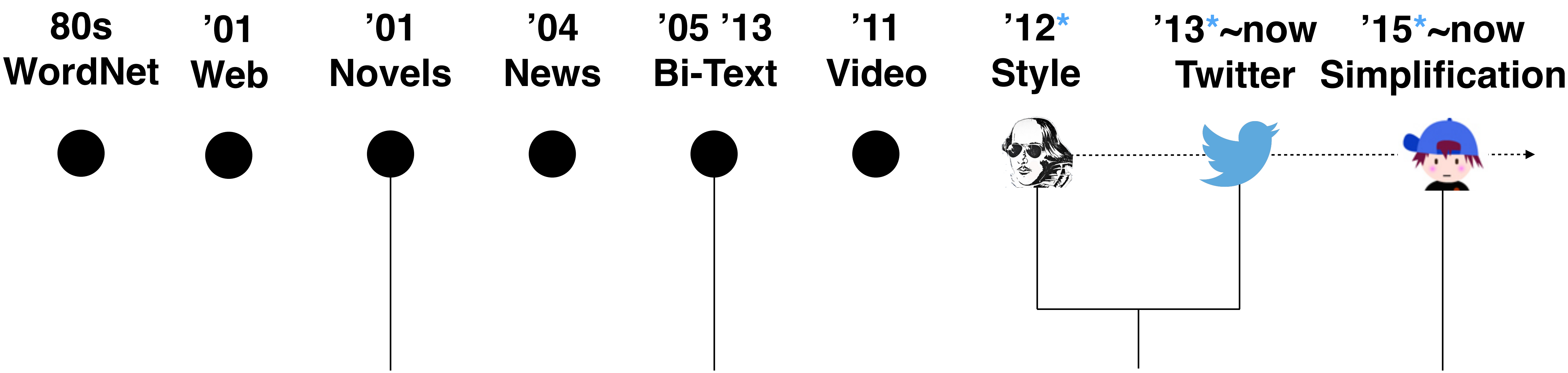


Stronger evidence can be much harder to find and score…

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms

Paraphrase Research

* my research



Regina Barzilay. “Information Fusion for Multidocument Summarization: Paraphrasing and Generation”
PhD Thesis (2003)

Chris Callison-Burch. “Paraphrasing and Translation”
PhD Thesis (2007)

Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations”
PhD Thesis (2014)
NSF CRII Award (2018)

Percy Liang. “Interactive Training of Semantic Parsers via Paraphrasing”
NSF CAREER Award (2016)

Part 1: Data & Models

Neural Network Models for …

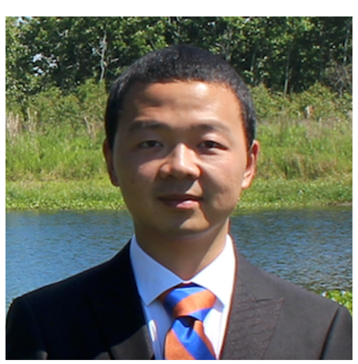
Paraphrase Identification

Semantic Textual Similarity

Natural Language Inference

Question Answering

Wuwei Lan and Wei Xu



(COLING 2018 - best paper award)

Paraphrase Identification
Semantic Textual Similarity
Natural Language Inference
Question Answering

Sentence Pair Modeling

The General Neural Framework

Type I: The Sentence Encoding-based Models

- Gated recurrent average network [Wieting and Gimpel, 2017]
- Directional self-attention network [Shen et al., 2017]
- **InferSent** BiLSTM with max-pooling [Conneau et al., 2017]
- Gumbel Tree-LSTM [Choi et al., 2017]
- **SSE** Shortcut-stacked BiLSTM [Nie and Bansal, 2017]
- and many others …

Type II: The Word Interaction-based Models

- **PWIM** Pairwise word interaction [He and Lin, 2016]
- Subword-based pairwise word interaction [Lan and Xu, 2018]
- Attention based CNN [Yin et al., 2016]
- **DecAtt** Decomposable attention [Parikh et al., 2017]
- **ESIM** Enhanced LSTM for NLI [Chen et al., 2017]
- and many others …

Motivation for this Work

		SNLI	MNLI	Quora	URL	PIT	STS14	WikiQA	TrecQA
Type I	**InferSent**	0.845	-	-	-	-	0.700	-	-
	SSE	0.860	0.746	-	-	-	-	-	-
Type II	**DecAtt**	0.863	-	0.865	-	-	-	-	-
	ESIM_seq	0.880	0.723	-	-	-	-	-	-
	ESIM_tree	0.878	-	-	-	-	-	-	-
	ESIM_seq+tree	0.886	-	-	-	-	-	-	-
	PWIM	-	-	-	0.749	0.667	0.767	0.709	0.759

- Previous systems only reported results on a few selected datasets.

Reproduced Results for Sentence Pair Modeling

		SNLI	MNLI	Quora	URL	PIT	STS14	WikiQA	TrecQA
Type I	**InferSent**	0.846	0.705	0.866	0.746	0.451	0.715	0.287	0.521
	SSE	0.855	0.740	**0.878**	0.650	0.422	0.378	0.624	0.628
Type II	**DecAtt**	0.856	0.719	0.865	0.652	0.430	0.317	0.603	0.660
	ESIM_seq	0.870	0.752	0.850	0.748	0.520	0.602	0.652	**0.771**
	ESIM_tree	0.864	0.736	0.755	0.740	0.447	0.493	0.618	0.698
	ESIM_seq+tree	**0.871**	**0.753**	0.854	0.759	0.538	0.589	0.647	0.749
	PWIM	0.822	0.722	0.834	**0.761**	**0.656**	**0.743**	**0.706**	0.739

- We filled in the blanks and systematically compared 7 models on 8 datasets.
- No model consistently performs well across all tasks!

4 Tasks and 8 Benchmark Datasets

Paraphrase Identification

paraphrase | non-paraphrase

Dataset: Quora (400k), URL (51k), PIT (16k)

Semantic Textual Similarity

score[0,5]

Dataset: STS14 (11k)

Natural Language Inference

entailment | neutral | contradiction

Dataset: SNLI (570k), MNLI (432k)

Question Answering

true | false

Dataset: WikiQA (12k), TrecQA (56k)

Example Datasets

Question Answering [1]

true

false

Q: How much is 1 tablespoon of water?

A: In Australia one tablespoon (measurement unit) is 20 mL

A: It is abbreviated as t, tb, tbs, tbsp, tblsp, or tblspn.

Paraphrase Identification [2]

paraphrase

non-paraphrase

CO2 levels haven't been this high for 3 to 5 million years.

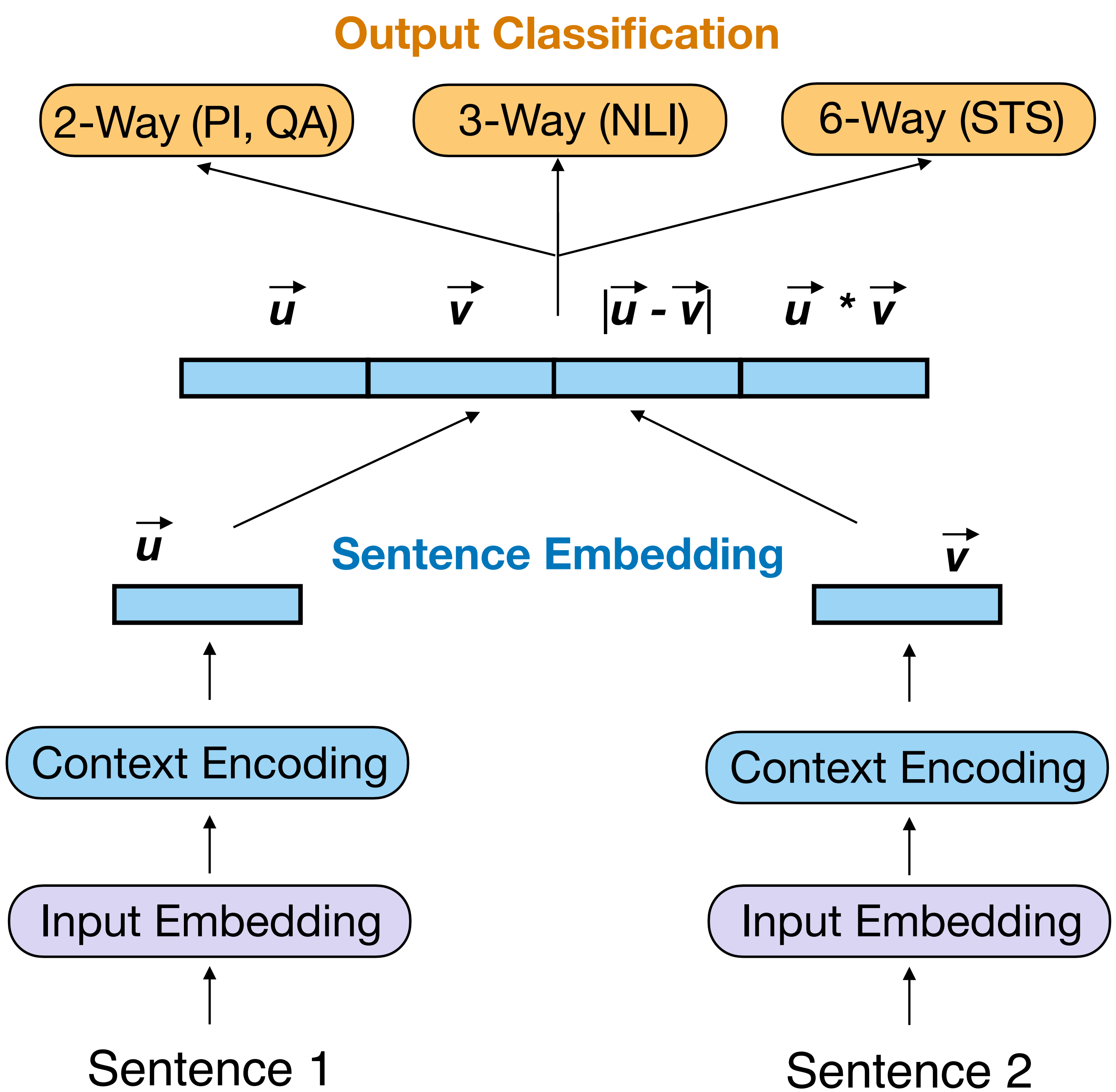
CO2 levels mark 'new era' in the world's changing climate.

First whole year over 400ppm. We are too complacent with this news .

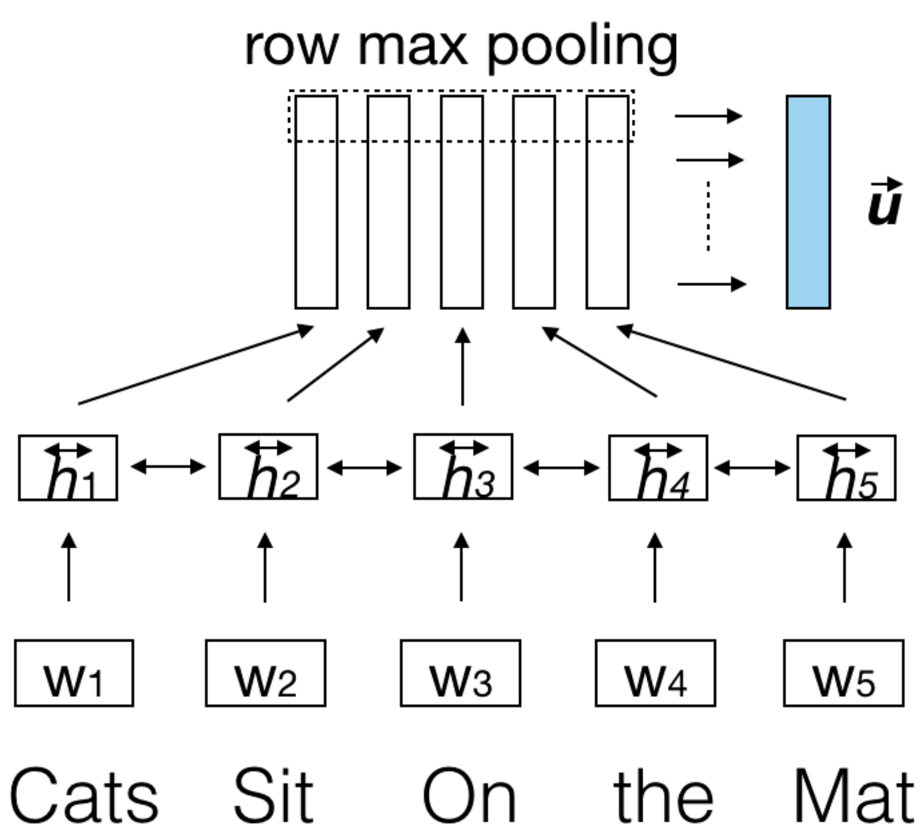
[1] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. (EMNLP 2015).

[2] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A Continuously Growing Dataset of Sentential Paraphrases (EMNLP 2017).

Type I: Sentence Encoding-based Models



Type I: Sentence Encoding-based Models

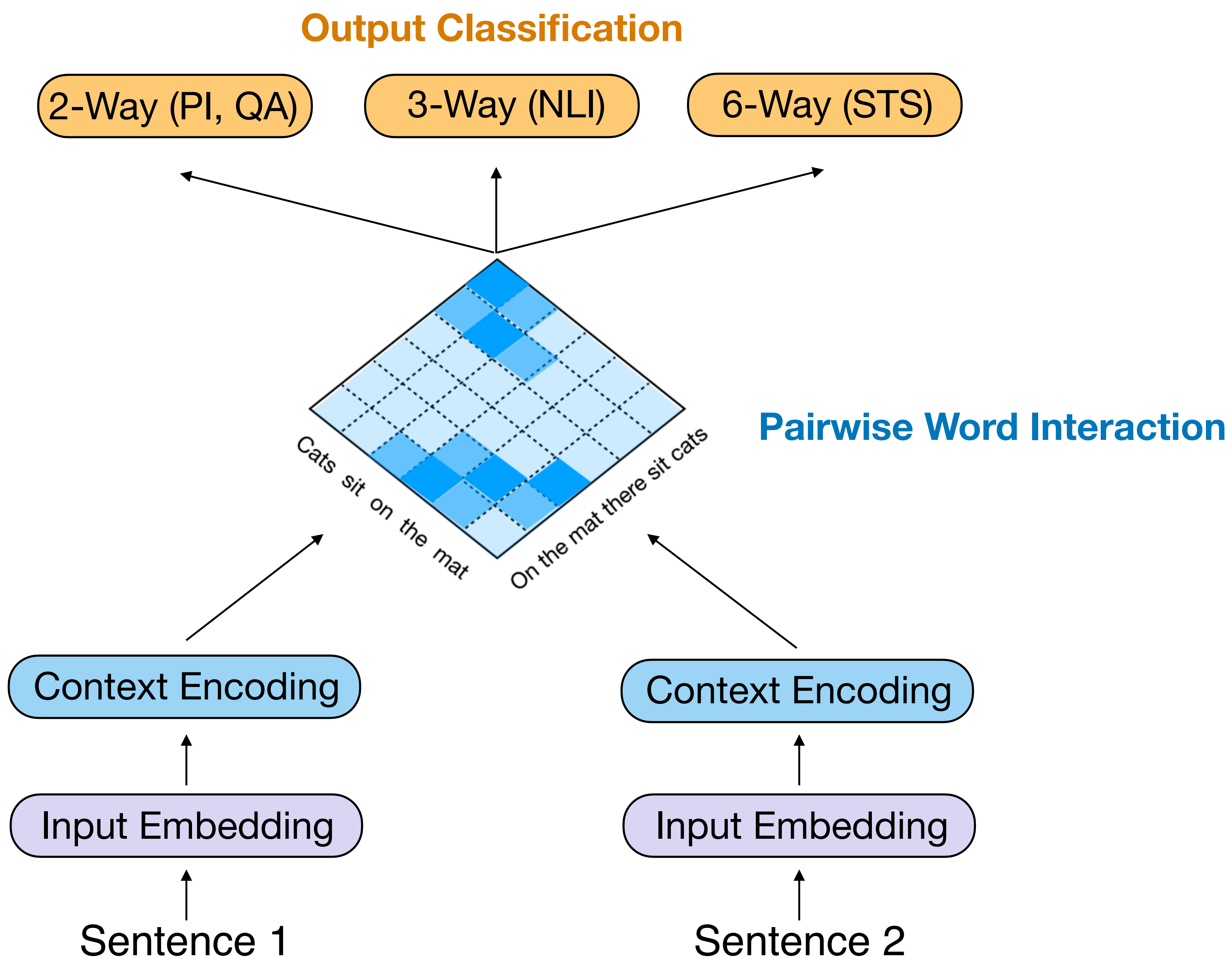


Sentence Embedding

InferSent: 1-layer Bi-LSTM.[3]
SSE: 3-layer Bi-LSTM with skip connection.[4]

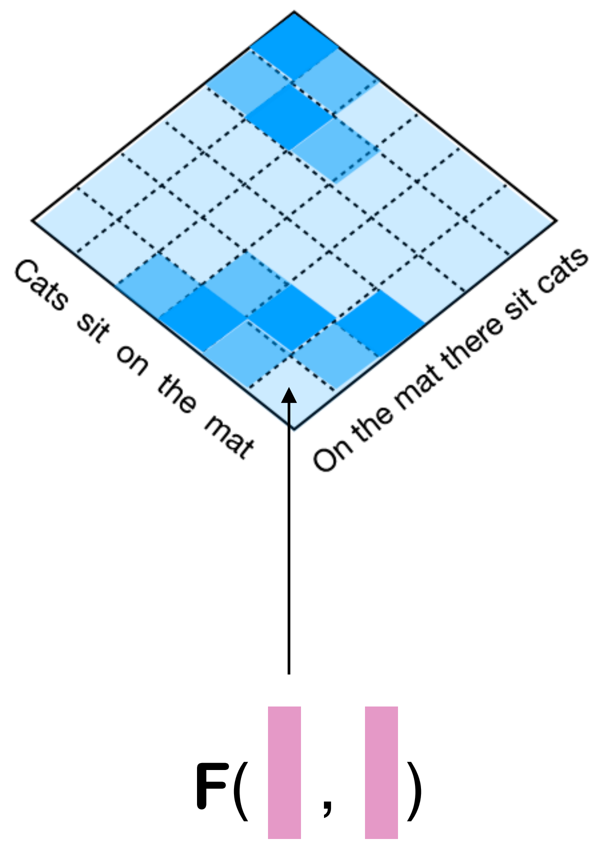
[3] Jihun Choi, Kang Min Yoo, and Sang-goo Lee: Unsupervised learning of task-specific tree structures with tree-LSTMs. (EMNLP 2017).
[4] Yixin Nie and Mohit Bansal. Shortcut-stacked sentence encoders for multi-domain inference. (RepEval 2017)

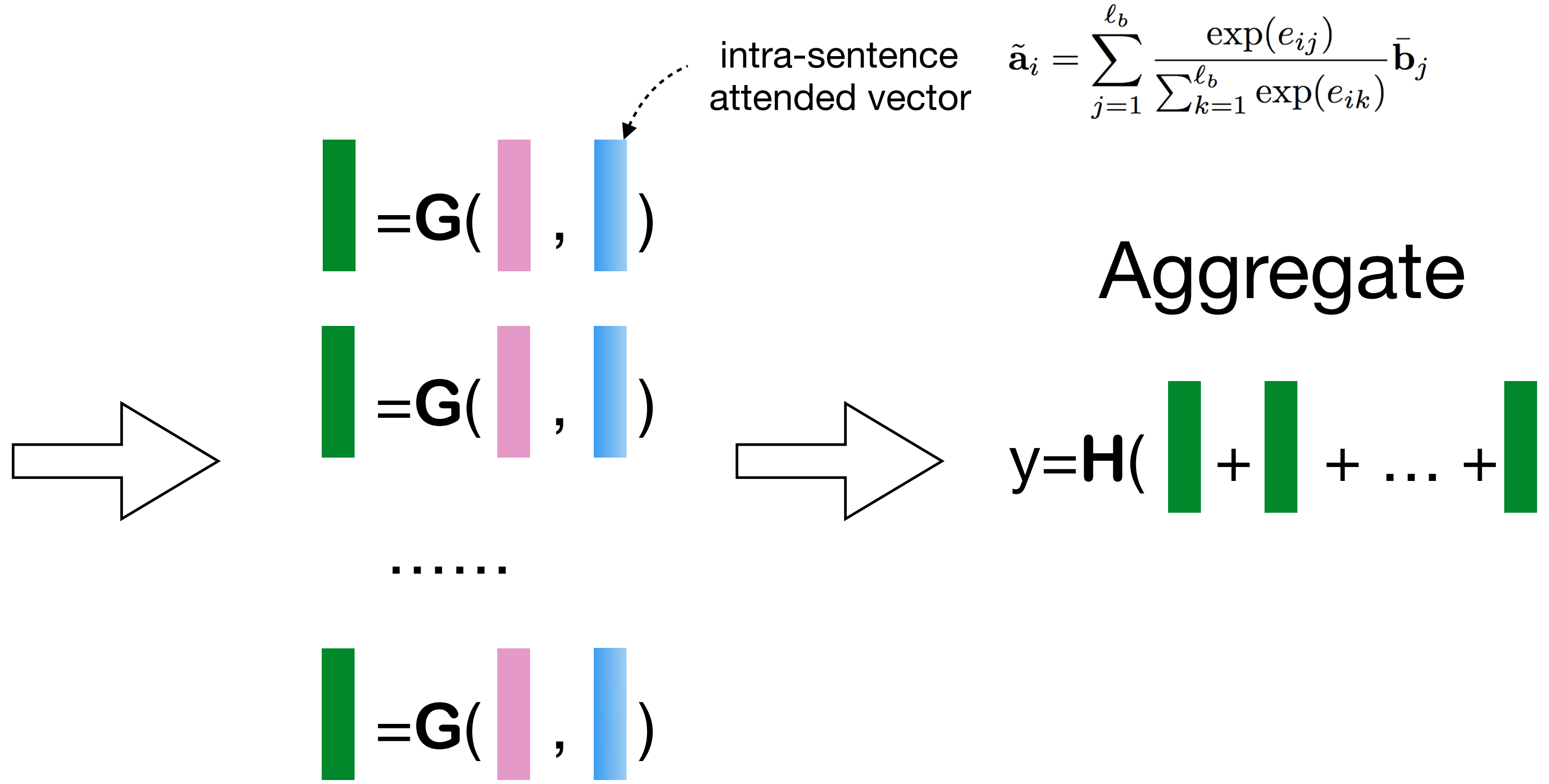
Type II: Word Interaction-based Models



- semantic relation between two sentences depends largely on aligned words/phrases

Pairwise Word Interaction





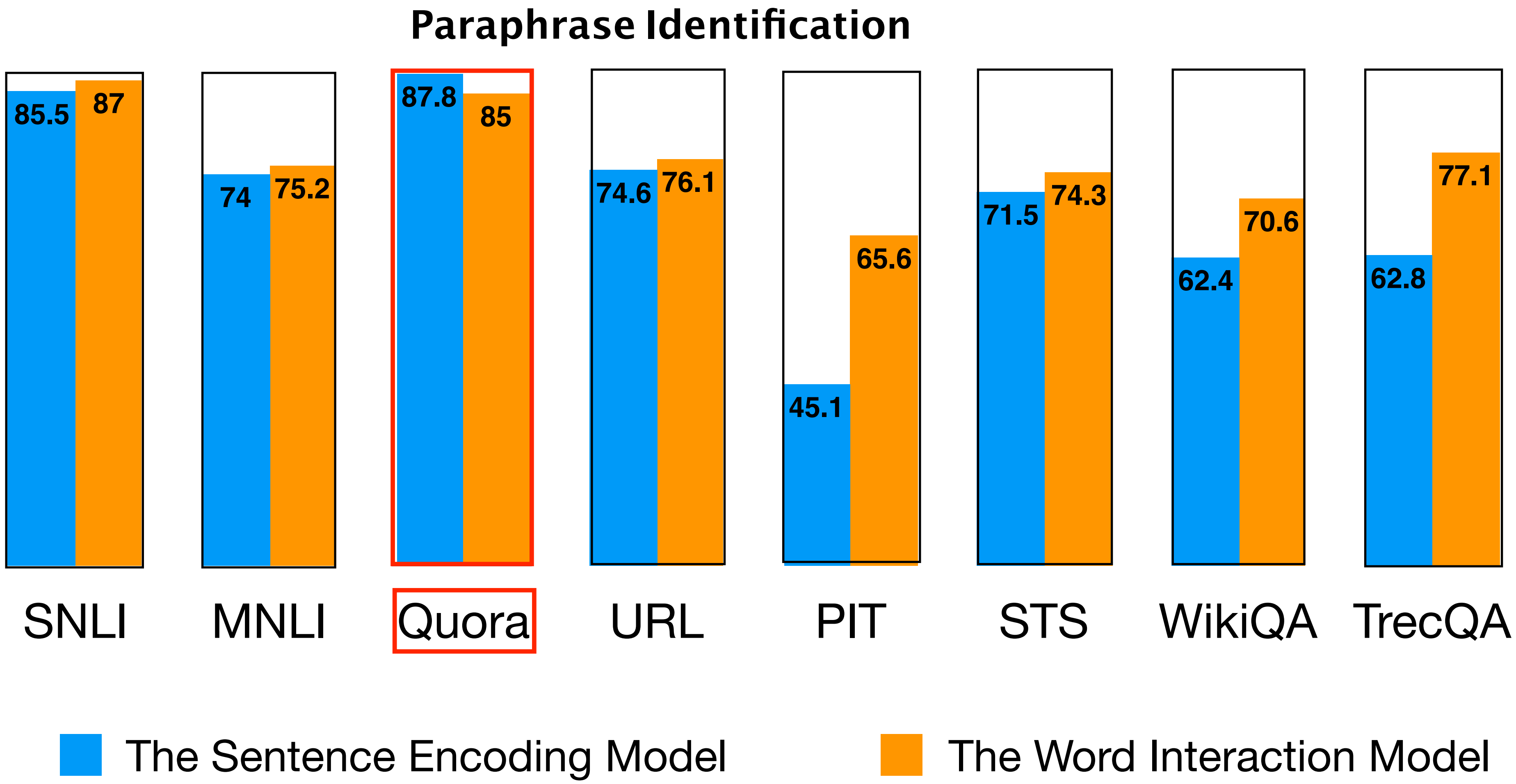
DecAtt[5]: **F** is dot product; **G, H** are feedforward networks.

ESIM[6]: more features in **G**(, , −, ⊙), and **G** is replaced with Bi-LSTM/Tree-LSTM.

PWIM[7]: **F** uses cosine, L2 and dot product; **G** (,) is “hard” attention; **H** is deep CNN.

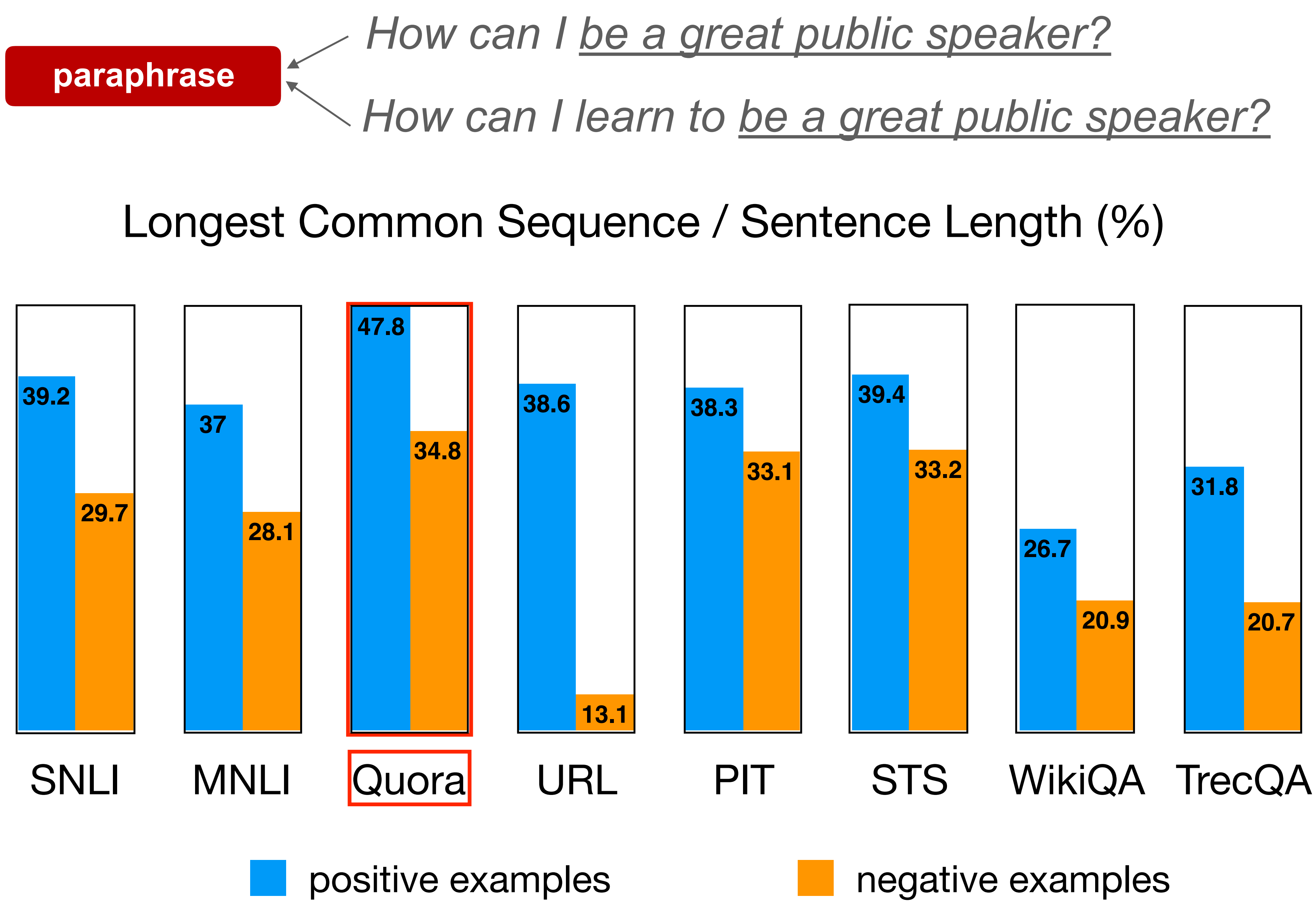
[5] Ankur Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkorei. A decomposable ¨ attention model for natural language inference. (EMNLP 2016)
[6] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. (ACL 2017)
[7] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. (NAACL 2016)

What Type of Model performs better ?



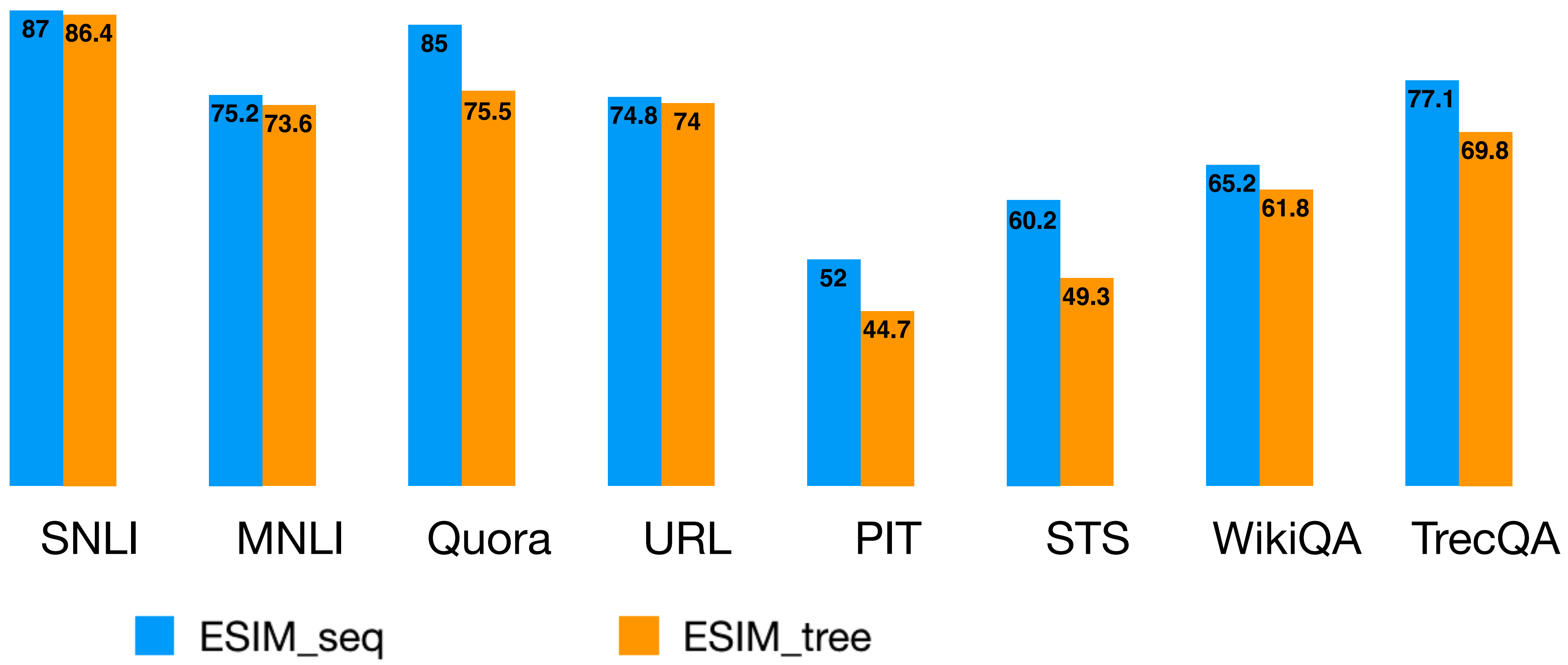
- Word Interaction-based Models perform much better (except Quora).

Why is Quora an exception ?



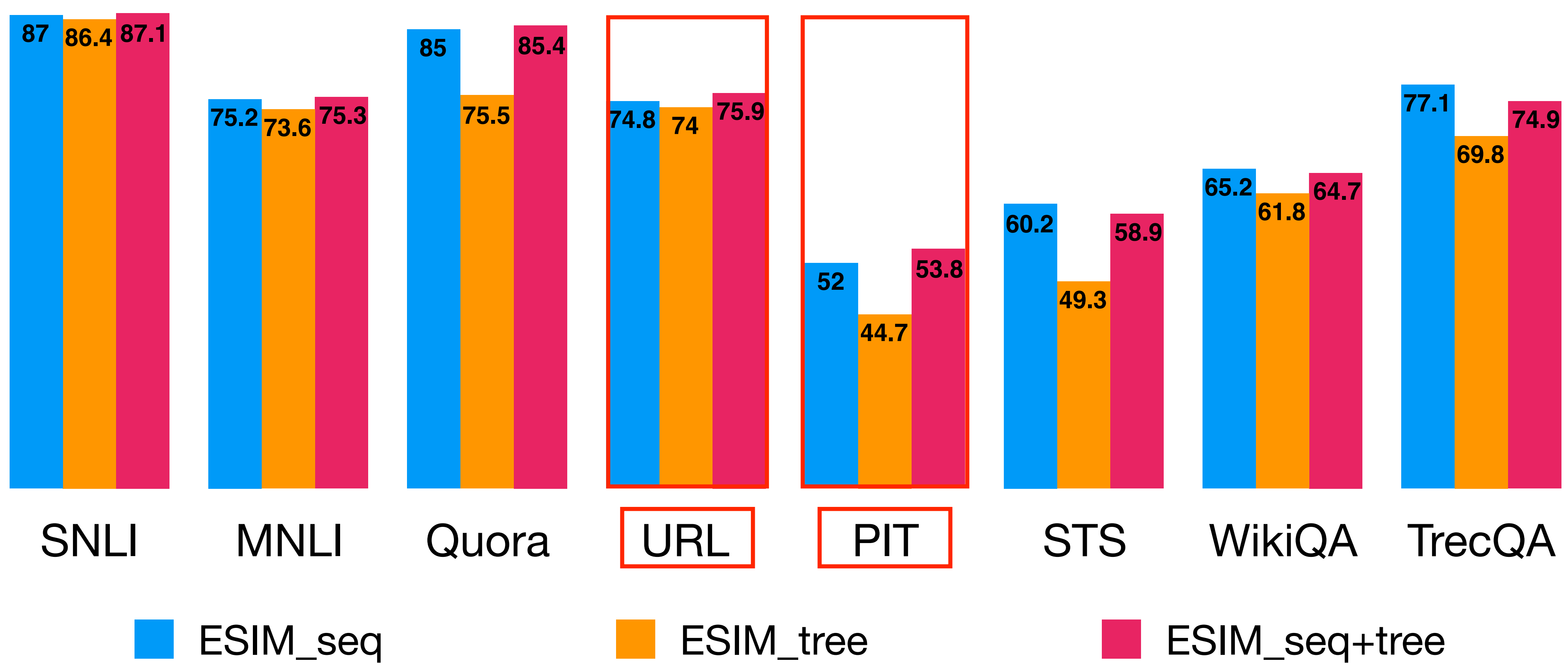
- Longer common sequences results in similar (RNN-based) sentence embeddings.

Bi-LSTM or Tree-LSTM?



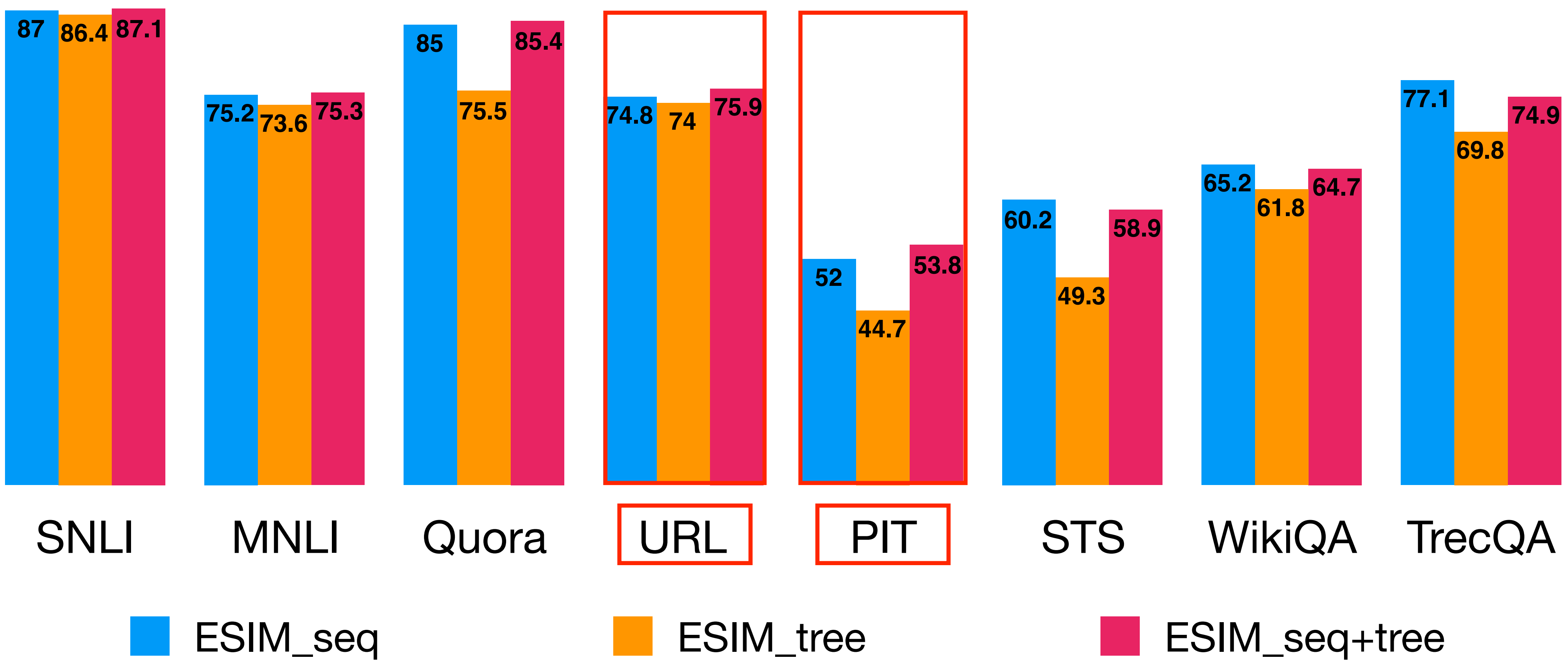
- ESIM_seq (Bi-LSTM) performs better than ESIM_tree (Tree-LSTM) on every dataset.

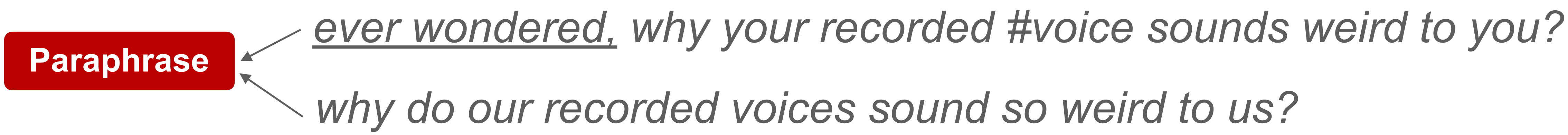
Bi-LSTM or Tree-LSTM?



- ESIM_seq (Bi-LSTM) performs better than ESIM_tree (Tree-LSTM) on every dataset.
- Adding Tree_LSTM (ESIM_seq+tree) helps on Twitter data (URL and PIT).

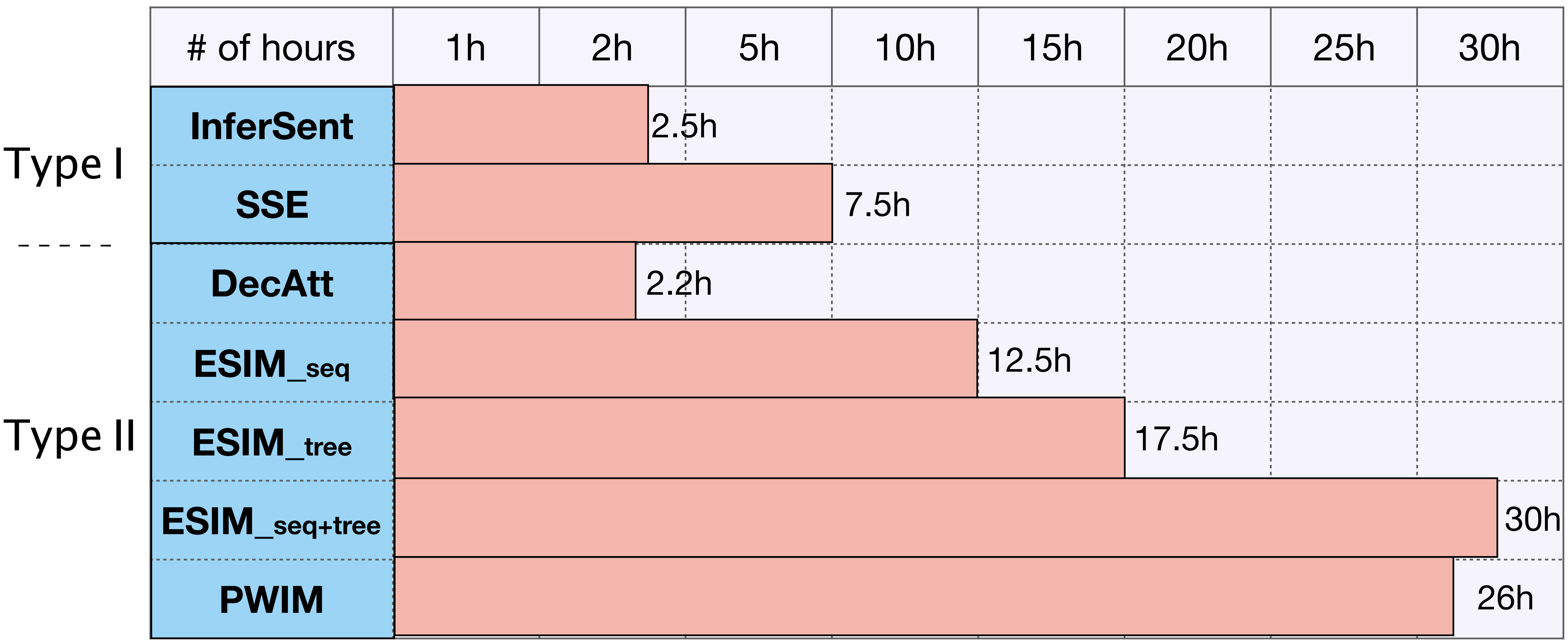
Why Tree-LSTM helps with Twitter data ?





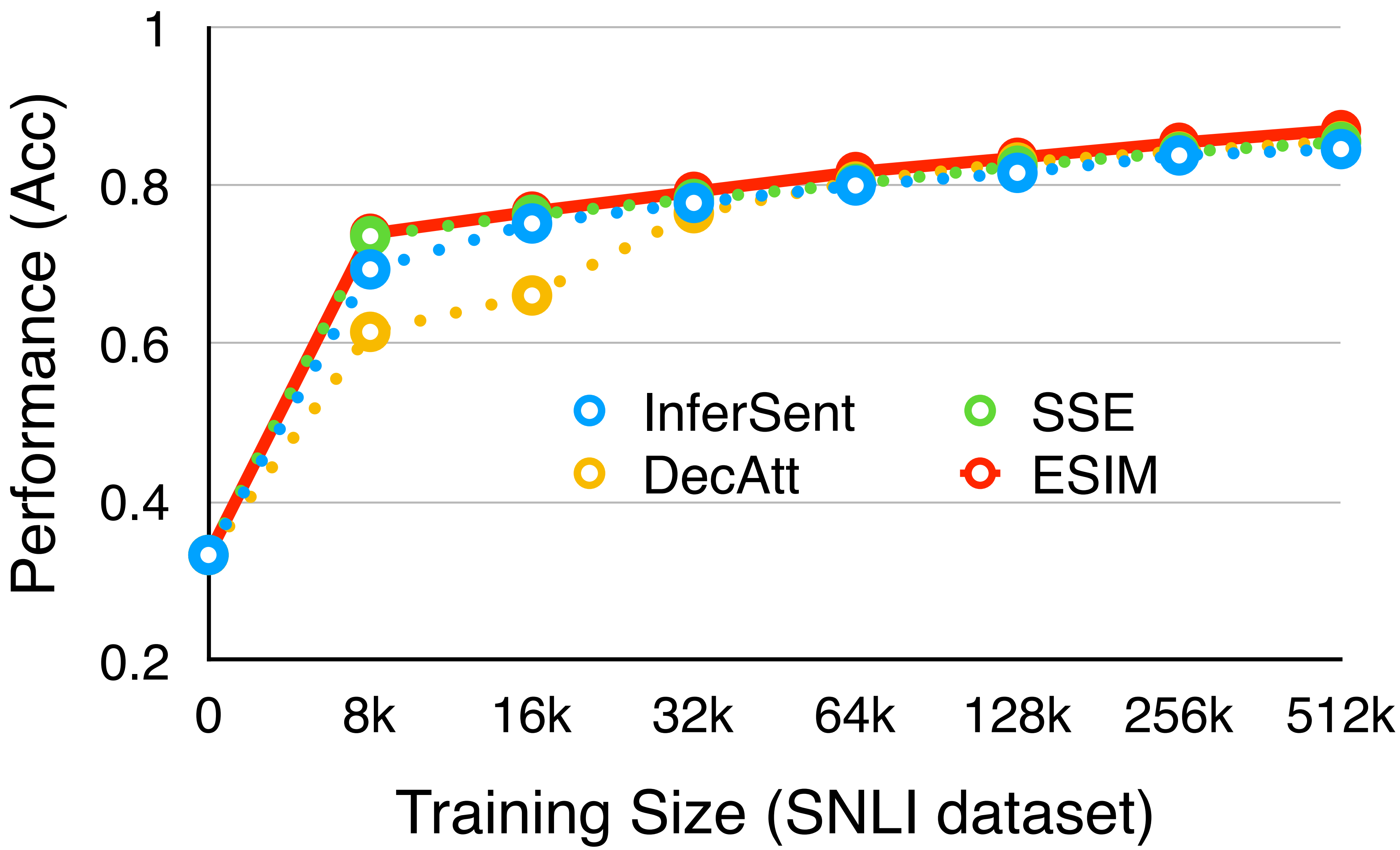
- Disruptive context can be put into less important position in Tree-LSTM.

Training Time on SNLI



- Training time comparison across different models on SNLI dataset (550k sent pairs).

Do we need more data?



- The learning curves are still increasing. More data can help!

We also need more **natural** data!

- Natural data — two sentences are written independently and have no label bias.

SNLI is large but contains data annotation artifacts. [8]

Twitter data contains natural paraphrases in large quantity, though can be noisy. [9]

more for future work!

paraphrase ← *Ezekiel Ansah is wearing real3D glasses with the lenses punched out*

Ezekiel Ansah wearing 3D glasses wout the lens

non-paraphrase ← *I wore the 3D glasses wout lenses before Ezekiel Ansah*

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data (NAACL 2018).
[9] Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, and Yangfeng Ji. Extracting Lexically Divergent Paraphrases from Twitter (TACL 2014).

Takeaways

- Systematic comparison of **5** representative models on **8** datasets
- **Large**, **clean**, and **more natural** data is needed for studying semantics!
- Code is available: **https://github.com/lanwuwei/SPM_toolkit**

Twitter as a powerful resource

thousands of users
talk about both big/micro events daily

a very broad range of paraphrases:
synonyms, misspellings, slang, acronyms and colloquialisms

Wei Xu, Chris Callison-Burch, Bill Dolan. SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (SemEval 2015)

Social Science

(also age & income)

Daniel Preotiuc, Wei Xu, Lyle Ungar. Discovering User Attribute Stylistic Differences via Paraphrasing. (AAAI 2016)

Part 2: Applications

A Word-Complexity Lexicon and **A Neural Readability Ranking Model** for Lexical Simplification (EMNLP 2018)

Mounica Maddela and Wei Xu

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 **<u>puree</u>** made of apples.*

OUTPUT: *Applesauce is a **<u>soft paste</u>**. It is made of apples.*

thick liquid

liquidized sauce

complex

Complex Word Identification - Substitution Generation - Substitution Ranking

A Large Word-complexity Lexicon

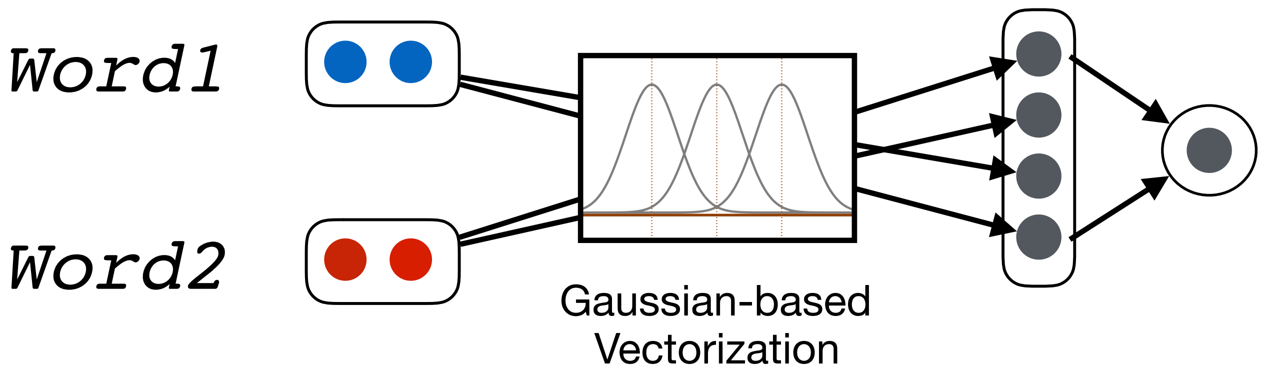
- 15,000 English words w/ human ratings

day	1.0
convenient	2.4
transmitted	3.2
cohort	4.3
assay	5.8

MIN 1 (simple)

MAX 6 (complex)

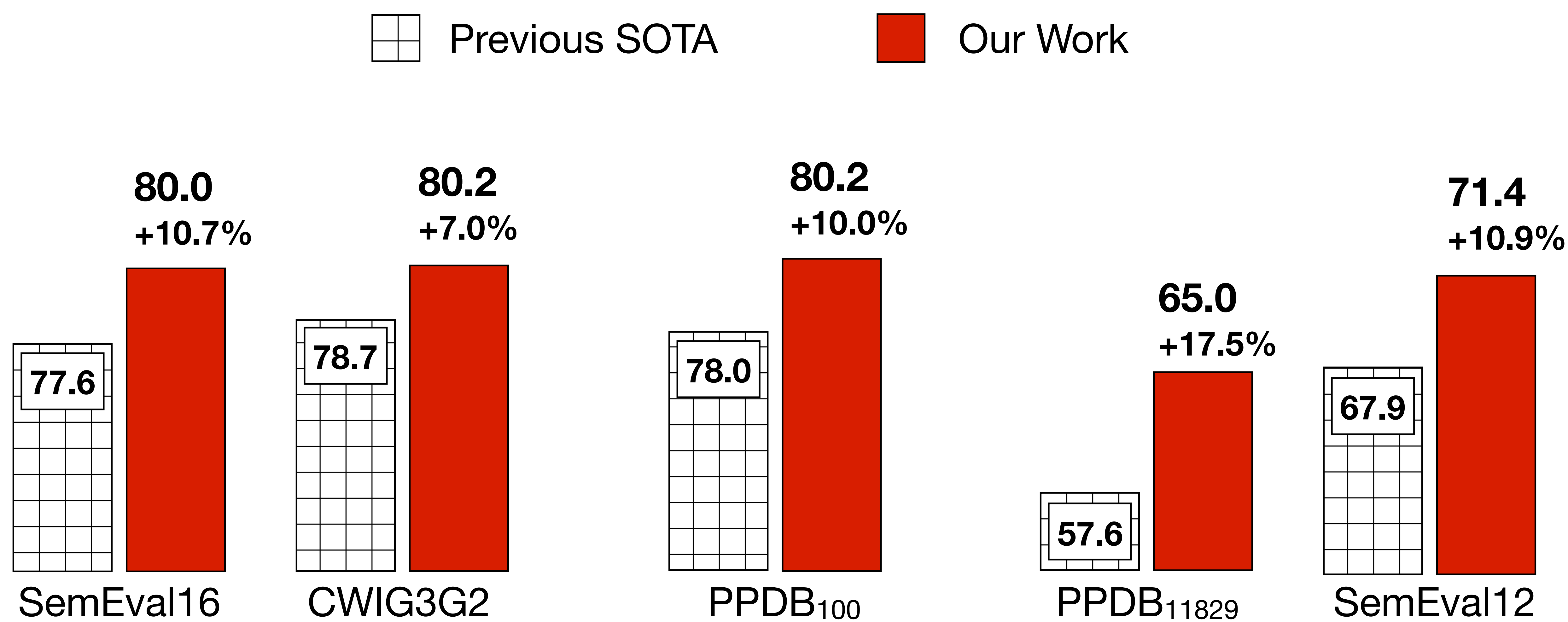
- predict relative complexity for any given words or phrases



P(*Word1* > *Word2*) = 0.91

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

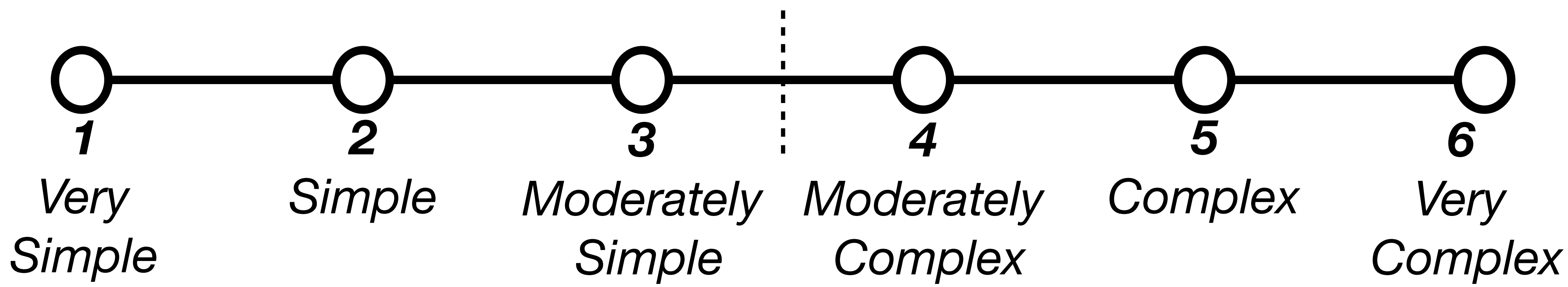
Assumption #2: more frequent words are simpler → **Wrong! (14% of time*)**

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

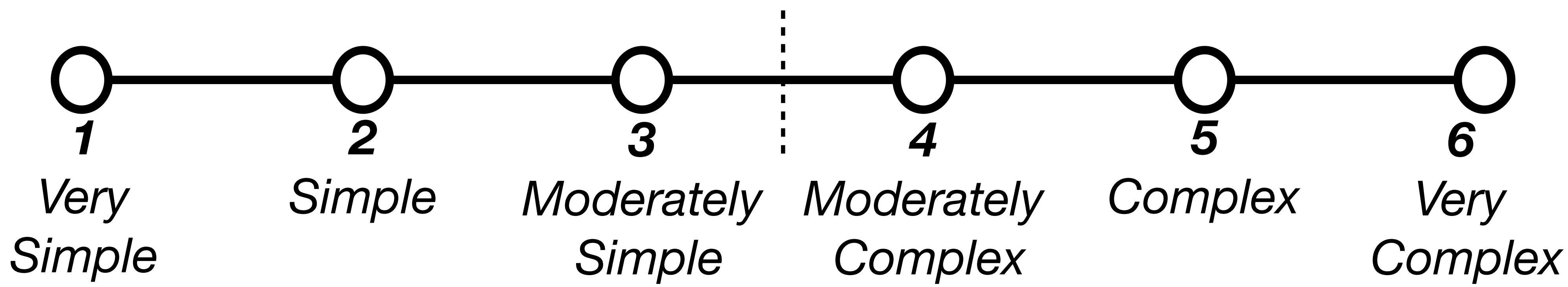
* based on 2272 lexical paraphrases sampled from PPDB

A Large Word-complexity Lexicon

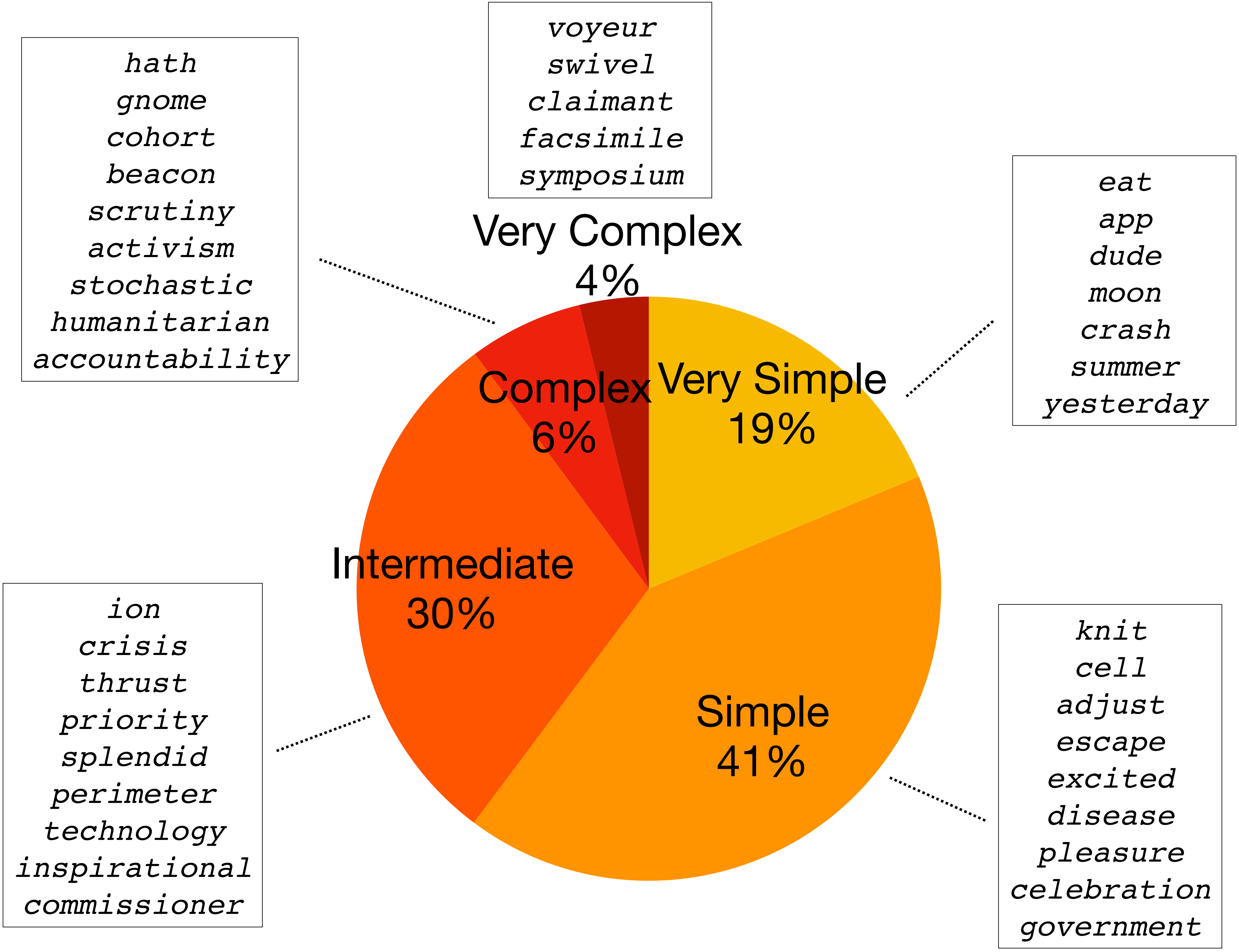
- 15,000 most frequent English words from Google 1T ngram corpus
- Rated on a 6-point Likert scale



- 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
Very Complex
4%
eat
app
dude
moon
crash
summer
yesterday
Complex
6%
Very Simple
19%
Intermediate
30%
Simple
41%
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
`muscles`	1.6	2	1	2	2	1
`pattern`	2.4	2	3	1	1	3
`educational`	3.2	3	3	3	3	4
`cortex`	4.2	4	4	4	4	5
`assay`	5.8	6	6	6	5	6

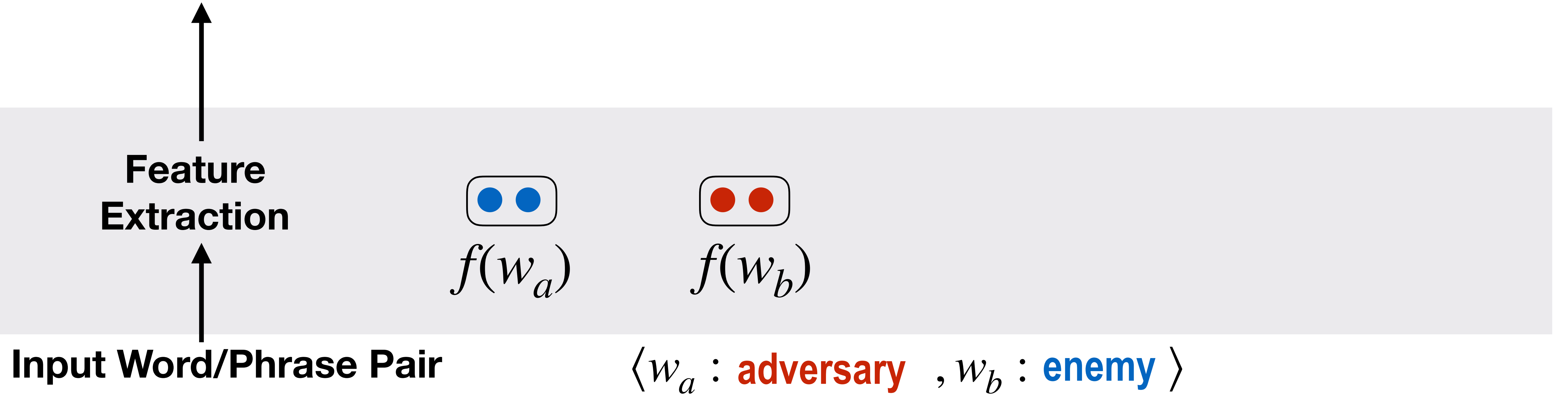
difference (one vs. rest)

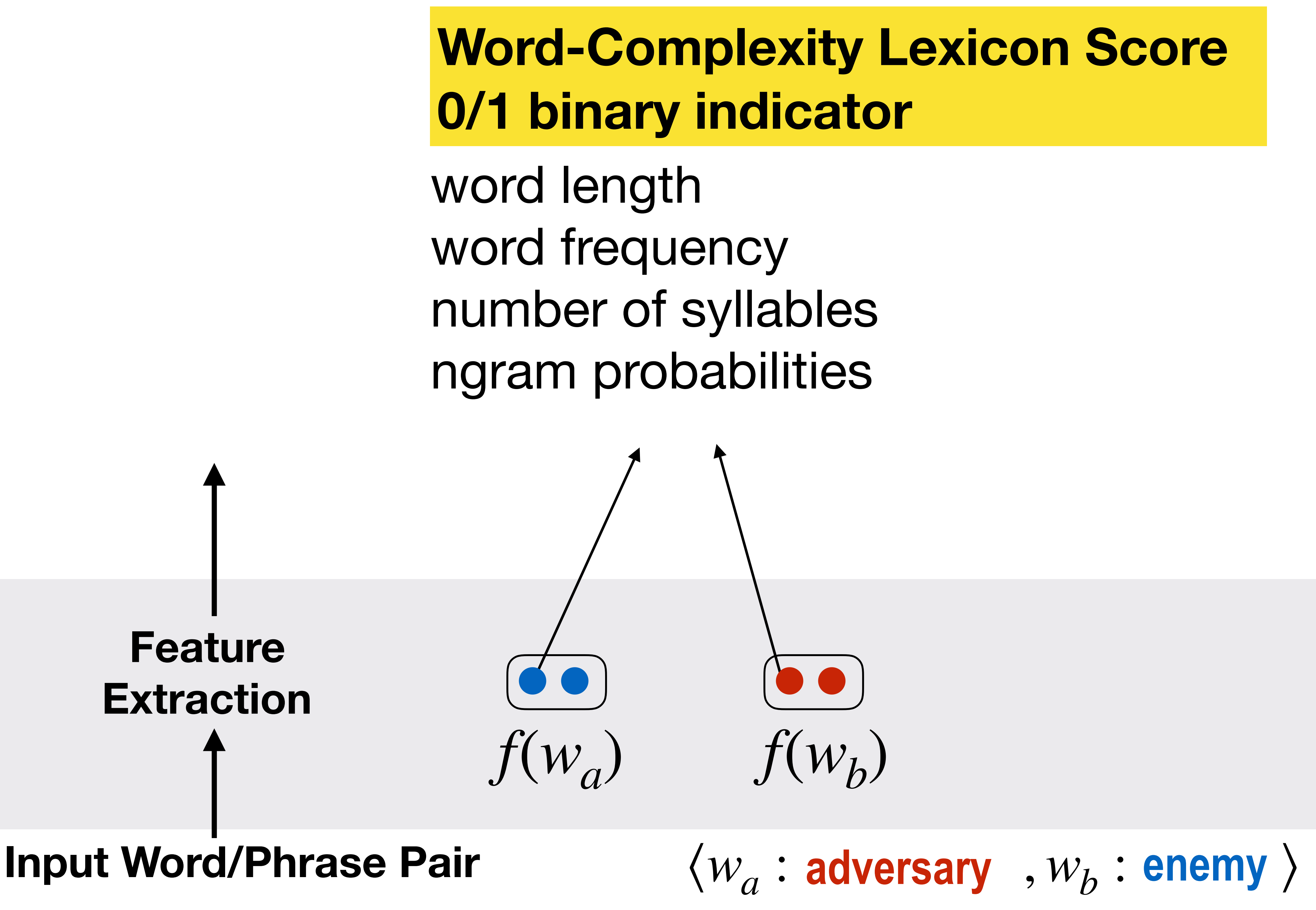
$<$ **0.5** for **47%** of annotations

$<$ **1.0** for **78%** of annotations

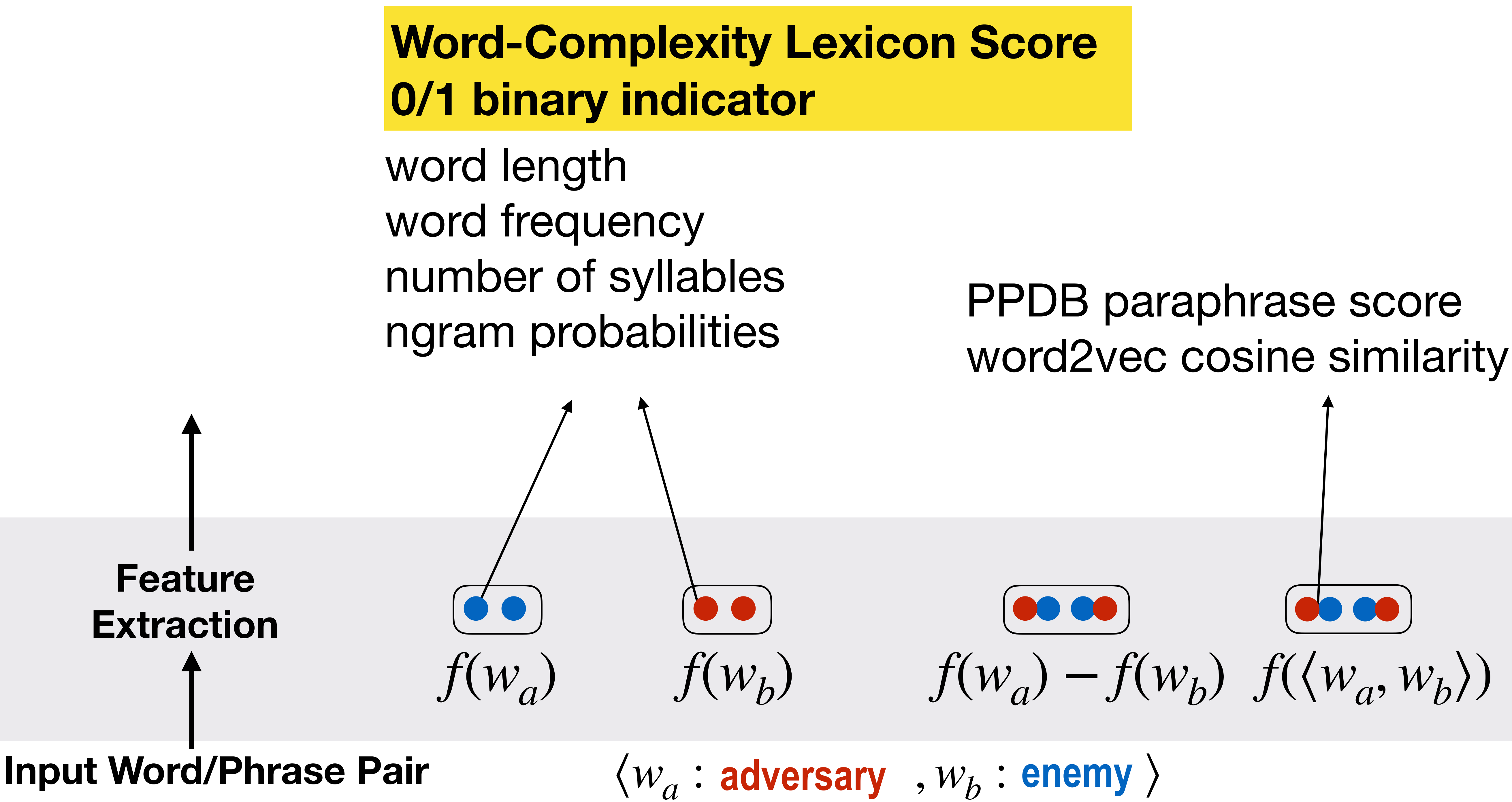
$<$ **1.5** for **93%** of annotations

A Pairwise Neural Ranking Model

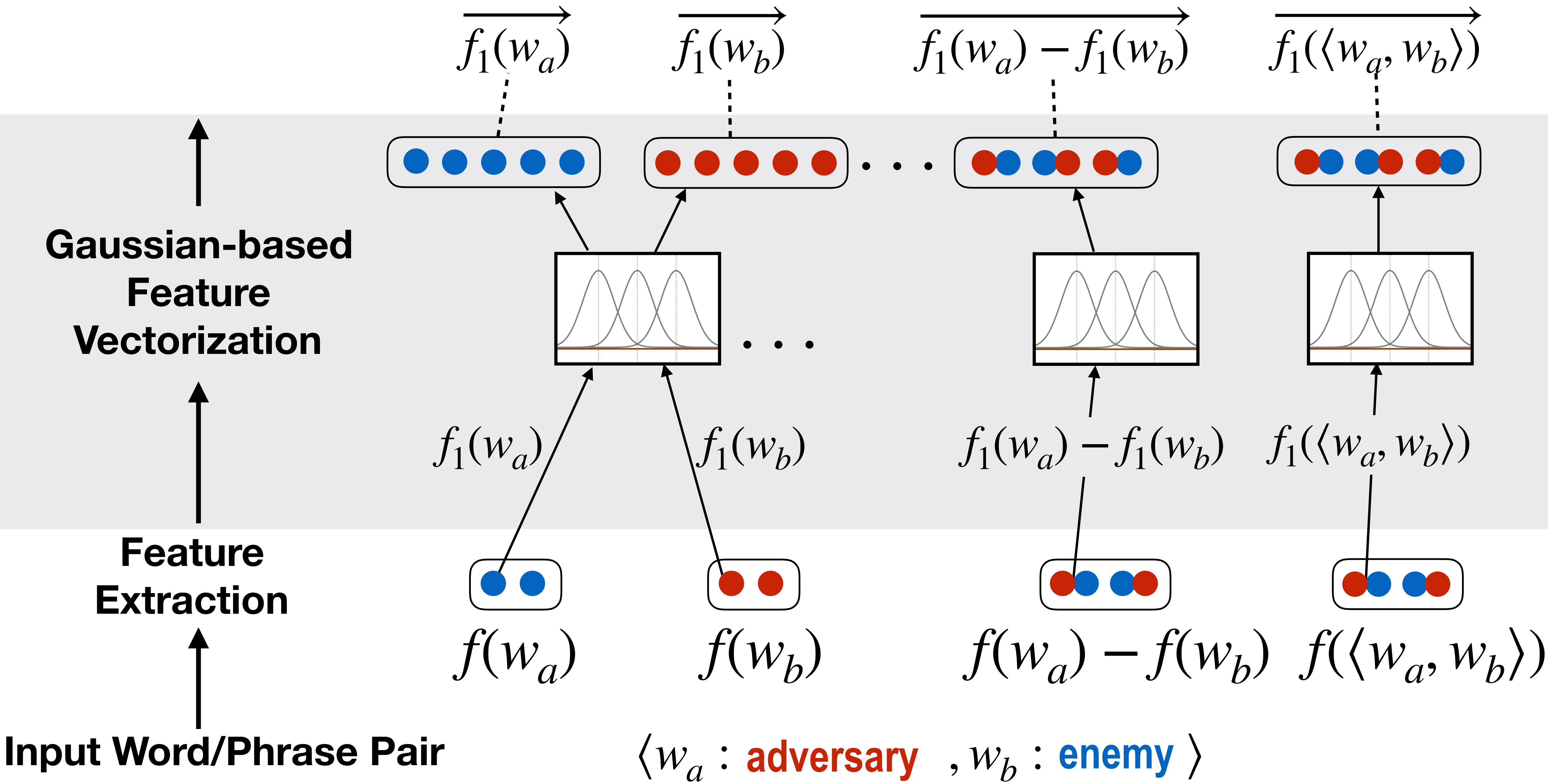

Feature Extraction
$f(w_a)$
$f(w_b)$
Input Word/Phrase Pair
$\langle w_a$: adversary , w_b : enemy $\rangle$



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
$\langle w_a$: adversary , w_b : enemy $\rangle$

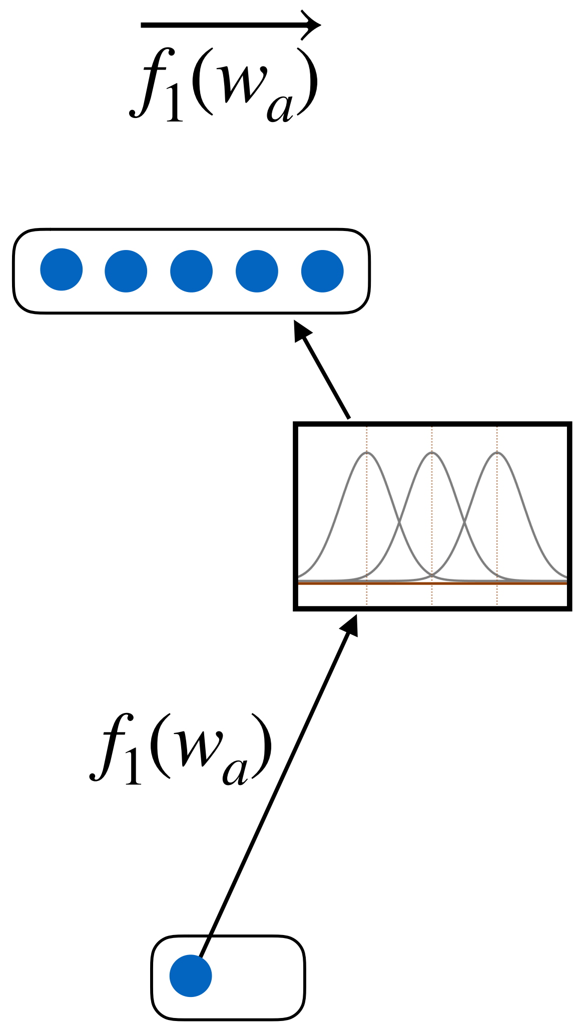


Word-Complexity Lexicon Score
0/1 binary indicator
word length
word frequency
number of syllables
ngram probabilities
PPDB paraphrase score
word2vec cosine similarity
Feature Extraction
$f(w_a)$
$f(w_b)$
$f(w_a) - f(w_b)$
$f(\langle w_a, w_b \rangle)$
Input Word/Phrase Pair
$\langle w_a$: adversary , w_b : enemy $\rangle$



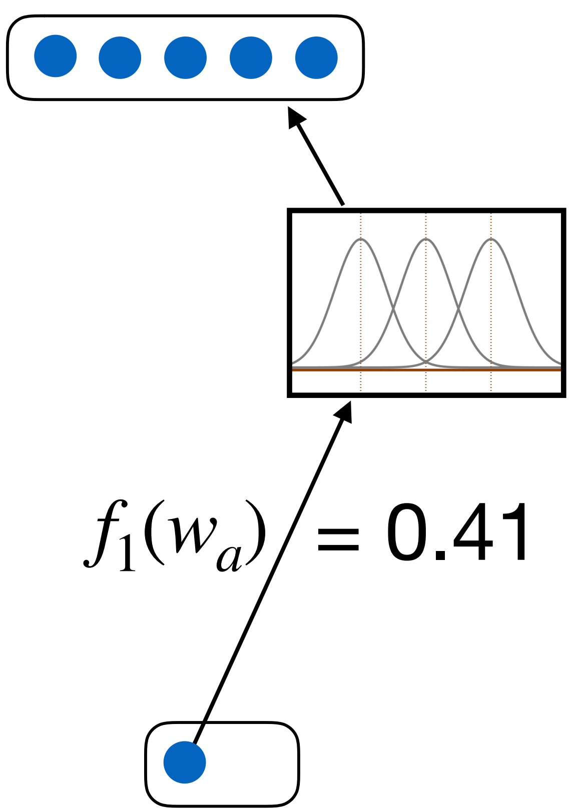
$\overrightarrow{f_1(w_a)}$
$\overrightarrow{f_1(w_b)}$
$\overrightarrow{f_1(w_a) - f_1(w_b)}$
$\overrightarrow{f_1(\langle w_a, w_b \rangle)}$
Gaussian-based Feature Vectorization
$f_1(w_a)$
$f_1(w_b)$
$f_1(w_a) - f_1(w_b)$
$f_1(\langle w_a, w_b \rangle)$
Feature Extraction
$f(w_a)$
$f(w_b)$
$f(w_a) - f(w_b)$
$f(\langle w_a, w_b \rangle)$
Input Word/Phrase Pair
$\langle w_a$: adversary , w_b : enemy $\rangle$

Gaussian-based Feature Vectorization



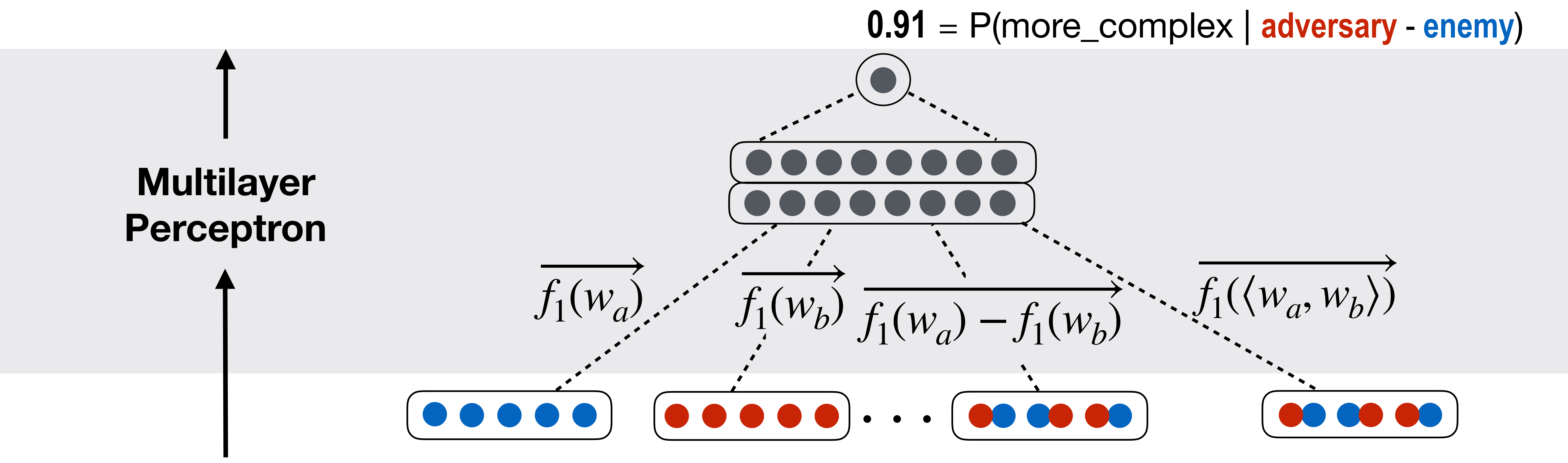
$\overrightarrow{f_1(w_a)}$ = [~0.0, **0.44**, **0.54**, ~0.02, ~0.0]

Gaussian-based Feature Vectorization



$$d_j(f) = e^{-\frac{(f-\mu_j)^2}{2\sigma^2}}$$

$f_1(w_a)$ = 0.41



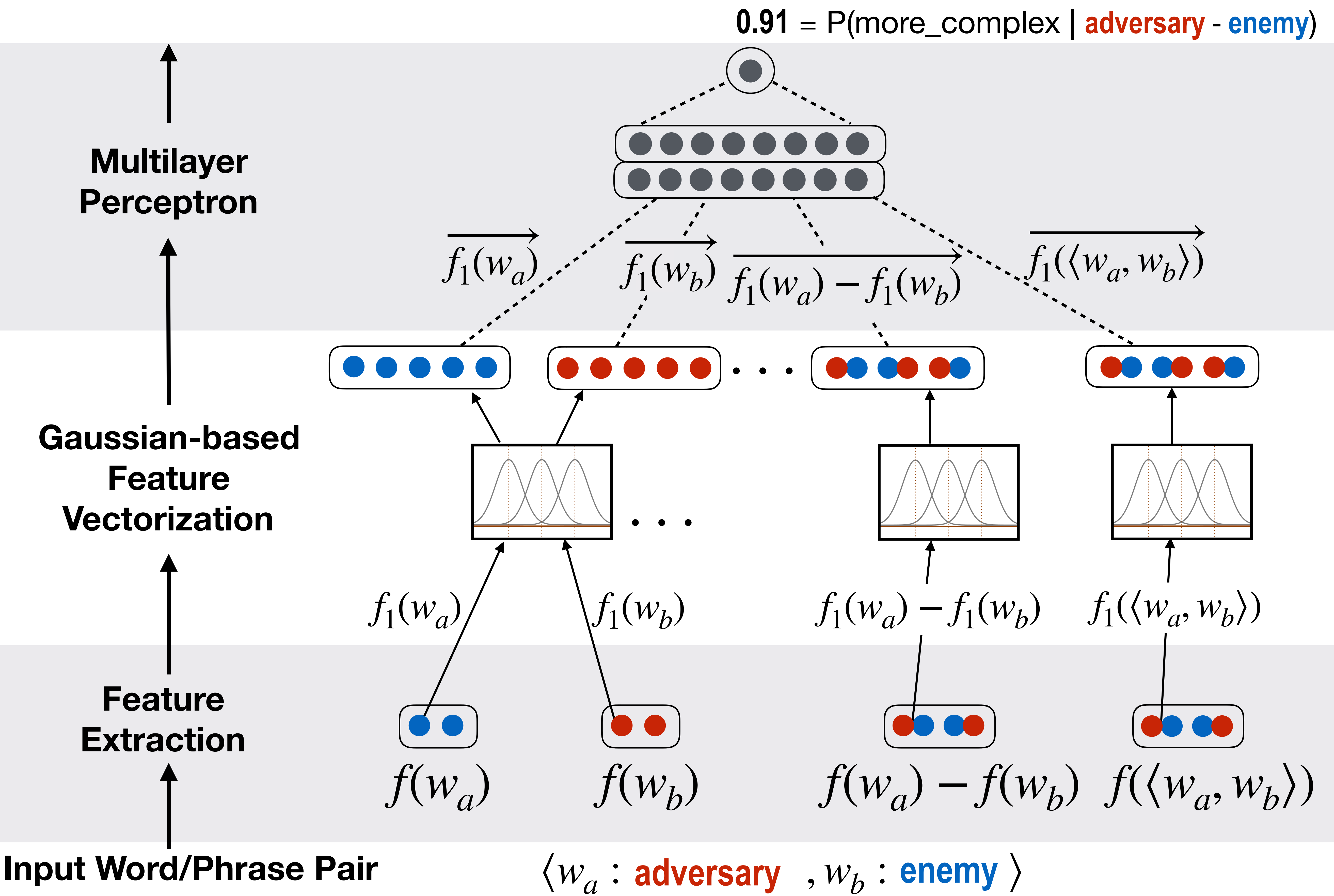
P > 0 $\Rightarrow$ w_a is more complex than w_b

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

| P | indicates complexity difference

$\langle w_a$: **adversary** , w_b : **enemy** $\rangle$

Neural Readability Ranking Model



Evaluation**

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

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

** 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 (+0.2)	0.679 (+0.002)
neural	Our Model + Lexicon + Gaussian	67.3* (+1.7)	0.714* (+0.035)

* 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 +0.002

+1.7 +0.035

* 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 (+0.2)	0.679 (+0.002)
neural	Our Model	65.4	0.682
neural	Our Model + Gaussian	66.6	0.702*
neural	Our Model + Lexicon + Gaussian	67.3* (+1.7)	0.714* (+0.035)

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

Evaluation - Error Analysis

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

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

SimplePPDB++

- 14.1 million paraphrase rules w/ improved complexity ranking scores

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

Takeaways

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

day	1.0
convenient	2.4
transmitted	3.2
cohort	4.3
assay	5.8

MIN 1 (simple)

MAX 6 (complex)

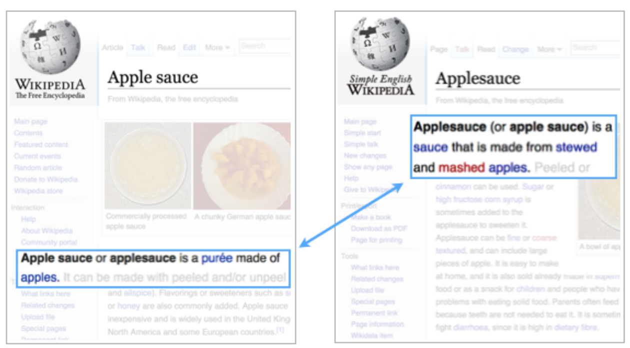
- PyTorch Code for the **Neural Ranking model** is also available!

  https://github.com/mounicam/lexical_simplification

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

Text Simplification

Parallel Wikipedia Corpus



(108k sentence pairs)

Phrase Table

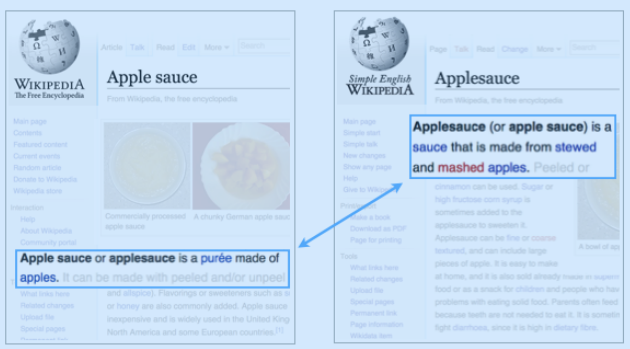
Statistical Machine Translation (often used as a blackbox)

standard approach since 2010

Wei Xu, Chris Callison-Burch, Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help (TACL 2015) .

Text Simplification

Parallel Wikipedia Corpus



(108k sentence pairs)

Phrase Table

Statistical Machine Translation (often used as a blackbox)

I showed that this setup is suboptimal, and why! *

*** my research: https://vimeo.com/150290363**

Wei Xu, Chris Callison-Burch, Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help (TACL 2015) .

Text Simplification

Feature Functions

PRO Tuning
(Hopkins & May, 2011)

$$
\begin{aligned}
g(i,j) > g(i,j') &\Leftrightarrow h_{\mathbf{w}}(i,j) > h_{\mathbf{w}}(i,j') \\
&\Leftrightarrow h_{\mathbf{w}}(i,j) - h_{\mathbf{w}}(i,j') > 0 \\
&\Leftrightarrow \mathbf{w} \cdot \mathbf{x}(i,j) - \mathbf{w} \cdot \mathbf{x}(i,j') > 0 \\
&\Leftrightarrow \mathbf{w} \cdot (\mathbf{x}(i,j) - \mathbf{x}(i,j')) > 0
\end{aligned}
$$

Large-scale Paraphrases
(lexical, phrasal, syntactic)

Objective Function

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{R})}$$

Tuning Data
(crowdsourced multi-references)

a state-of-the-art sentence simplification system

(also 1st evaluation metric that correlates with human judgement)

Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. Optimizing Statistical Machine Translation for Simplification (TACL 2016)

Text Simplification

Feature Functions

Large-scale Paraphrases
(lexical, phrasal, syntactic)

PRO Tuning
(Hopkins & May, 2011)

$$\begin{aligned} g(i,j) > g(i,j') &\Leftrightarrow h_{\mathbf{w}}(i,j) > h_{\mathbf{w}}(i,j') \\ &\Leftrightarrow h_{\mathbf{w}}(i,j) - h_{\mathbf{w}}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot \mathbf{x}(i,j) - \mathbf{w} \cdot \mathbf{x}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot (\mathbf{x}(i,j) - \mathbf{x}(i,j')) > 0 \end{aligned}$$

Tuning Data
(crowdsourced multi-references)

Objective Function

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \bar{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \bar{R})}$$

It made possible to train neural generation models.

Our datasets, evaluation metric, and system are now commonly adopted in the simplification research.

Other Applications

Education

- **Text Simplification** & Reading Assistant
- Auto Correction & Writing Assistant

(Xu et al. TACL'15)
(Xu et al. TACL'16)
(Maddela, Xu EMNLP'18)
(Xu et al. EMNLP'11)
(Xu et al. COLING'12)
(Xu et al. BUCC'13)

Social Media / News

- **Paraphrase Identification & Semantics**
- Information Extraction & Summarization
- Computational Social Science

(Xu et al. TACL'14)
(Xu et al. SemEval'15)
(Lan, Qiu, He, Xu EMNLP'17)
(Lan, Xu NAACL'18)
(Lan, Xu COLING'18)
(Xu et al. ACL'06)
(Xu et al. ACL'13)
(Tabassum, Ritter, Xu EMNLP'16)
(Kulkaurni, Xu, Ritter, Machiraju NAACL'18)
(Preotiuc, Xu, Ungar AAAI'16)

Our work is sponsored by:

NSF

DARPA

Ohio Supercomputer Center

criteo

figure eight

NVIDIA

Follow @cocoweixu

THE OHIO STATE UNIVERSITY

- Located in Columbus, Ohio (14th biggest city in US)
- 7 professors and ~20 phd students in NLP (CS and Linguists)

THE OHIO STATE UNIVERSITY

Wei Xu

Alan Ritter

Eric Fosler-Lussier

Marie de Marneffe

Micha Elsner

Mike White

William Schuler

I am grateful

thank u 4 ur time

thanku

Thank You

gramercies

thanks a lot

thanking you

appreciate it

gratitude

thx

3x

tyvm

thanks

thank you very much

say thanks

thnx

thanks a ton

I can no other answer make but thanks,
And thanks, and ever thanks.

wawwww thankkkkkkkkkkk you alottttttttttttt!

Follow @cocoweixu

all software and data are available on my homepage:
http://web.cse.ohio-state.edu/~weixu/