



WiC Challenge

Jose Camacho-Collados

@SemDeep-5, IJCAI 2019
Macau, 12 August

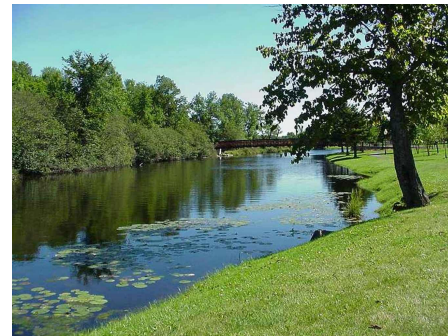


Word-in-Context: Motivation

*He withdrew money from the **bank**.*

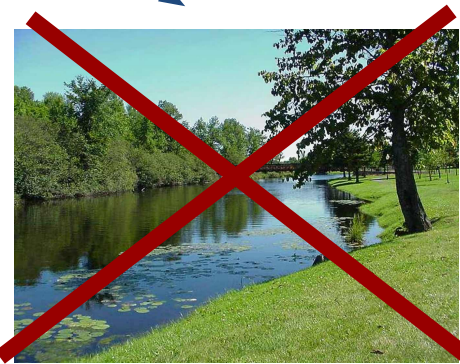
Word-in-Context: Motivation

*He withdrew money from the **bank**.*



Word-in-Context: Motivation

*He withdrew money from the **bank**.*



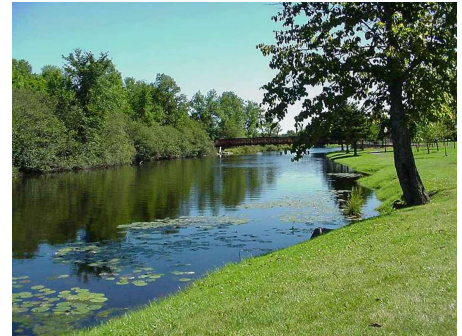


Word-in-Context: Motivation

*I want to sit by the **bank** of the river,
in the shade of the evergreen tree.*

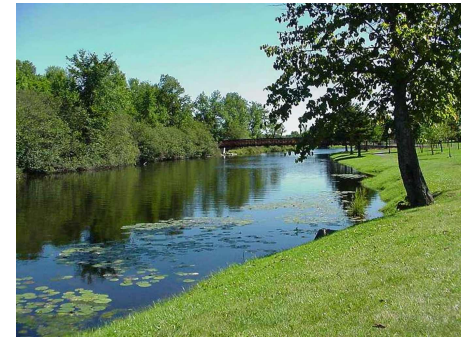
Word-in-Context: Motivation

*I want to sit by the **bank** of the river,
in the shade of the evergreen tree.*



Word-in-Context: Motivation

*I want to sit by the **bank** of the river,
in the shade of the evergreen tree.*





Word-in-Context: Motivation (at large)

Benchmark to test an important property (i.e. ambiguity) of **human language understanding** in machine (deep) learning models.

Part of a **wider effort to test different linguistic phenomena** (e.g. language inference, common sense, question answering, co-reference, etc.).

All these tasks under a single **general-purpose language understanding benchmark**:





Word-in-Context: What is it?

It is a task to evaluate context-sensitive representations of meaning (e.g. sense/contextualized embeddings, WSD systems, etc.)

Why? Words are ambiguous, and there is no suitable benchmark to test the dynamic nature of words' semantics.

For this challenge we proposed a **dataset for English**, based on:

[WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations](#)

M.T. Pilehvar and J. Camacho-Collados, Proc. of NAACL 2019 (Minneapolis, USA).



Word-in-Context: The task

Label	Target	Context-1	Context-2
F	bed	There's a lot of trash on the <u>bed</u> of the river	I keep a glass of water next to my <u>bed</u> when I sleep
F	land	The pilot managed to <u>land</u> the airplane safely	The enemy <u>landed</u> several of our aircrafts
F	justify	<u>Justify</u> the margins	The end <u>justifies</u> the means
T	beat	We <u>beat</u> the competition	Agassi <u>beat</u> Becker in the tennis championship
T	air	<u>Air</u> pollution	Open a window and let in some <u>air</u>
T	window	The expanded <u>window</u> will give us time to catch the thieves	You have a two-hour <u>window</u> of clear weather to finish working on the lawn



Word-in-Context: Main features

1. It is suitable for evaluating a wide range of techniques, including **contextualized word and sense representation** and **word sense disambiguation**.
2. It is framed as a **binary classification dataset**, in which identical words are paired with each other (in different contexts).
3. It is constructed using **high quality annotations** curated by experts.



Word-in-Context: The dataset

Contextual sentences in WiC were extracted from example usages provided for words in three lexical resources: **WordNet**, **VerbNet** and **Wiktionary**.

We used **WordNet** as the core resource, exploiting **BabelNet**'s mappings as a bridge between the resources.

Examples were compiled **semi-automatically**. **Pruning** and **manual verification** was performed as postprocessing.



Word-in-Context: Statistics

Split	Instances	Nouns	Verbs	Unique words
Training	5,428	49%	51%	1,256
Dev	638	62%	38%	599
Test	1,400	59%	41%	1,184

Statistics of different splits of WiC



WiC Challenge: Details of the competition

It was run through **March and April**.

Participants had ~ **three weeks** since release of test data.

Over ten teams submitted results, **seven were officially considered** (not all wrote task description paper here).



WiC Challenge: Results

Participants

Team	Accuracy (best)
w4ngatang	68.36
dloureiro	67.71
aina.gari	66.71
terachang	64.64
AlanAnsell	61.21
nishnik	55.43
gdls	51.93

Baselines

Baseline	Accuracy
<i>BERT-large</i>	65.5
<i>Context2vec</i>	59.3
<i>DeConf</i>	58.7
<i>SW2V</i>	58.1
<i>Elmo-3</i>	56.5
<i>JBT</i>	53.6
<i>Random</i>	50.0



Presentations

LIAAD at SemDeep-5 Challenge: Word-in-Context (WiC)

Daniel Loureiro and Alípio Mário Jorge

[\(next presentation\)](#)

An ELMo-inspired approach to SemDeep-5's Word-in-Context task

Alan Ansell, Felipe Bravo-Marquez and Bernhard Pfahringer

[\(presentation at 14:50\)](#)

LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC Challenge: Context Representations for Word Usage Similarity Estimation

Aina Garí Soler, Marianna Apidianaki and Alexandre Allauzen

[\(I will briefly present this now\)](#)

LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC Challenge: Context Representations for Word Usage Similarity Estimation

Aina Garí Soler¹, Marianna Apidianaki^{1,2}, Alexandre Allauzen¹



Features used

Combined **cosine similarities** from (different layers of) contextualized representations:

- SIF (Arora et al., 2017)
- Context2vec (Melamud et al., 2016)
- ELMo (Peters et al., 2018)
- BERT (Devlin et al., 2018)
- USE (Cert et al., 2018)

Automatically annotated substitutes:
(context2vec-based substitution)

- Proportion of common substitutes
- GAP score (Kishida, 2005)
- Substitute cosine similarity



Logistic regression classifier

Training Data Augmentation

4,018 pairs from Concepts in Context (ColnCo) corpus (Kremer et al., 2014) with **manual substitutes**

*The **man** closed his eyes.*

fellow
gentleman
guy
person

*This **man** is here in search of his lover (...)*

fellow
guy
male
person

SAME (75% overlap)

*I hopped out, **ran** after him, and reached him (...)*

chase trot
speed rush
go sprint
jog

*(...) effort to hire producers (...) to **run** the studio*

DIFF (0% overlap)

direct
film
manage
operate

Results and Analysis

- Best approach (**66.71**): combining cosines from **BERT** (av 4), **USE**, and **ELMo** (top, cw=2), **without data augmentation** and **without substitutes**

➔ *Why is ColnCo data not helping?*

- It contains clear-cut distinctions and doesn't work for **highly related but distinct (F) senses** in WiC:



Construction is underway on the new bridge (process)

The engineer marveled at his *construction* (result)

F




WiC Leaderboard

Contextualised word embeddings	Implementation	Accuracy %
BERT-large	Wang et al (2019)	68.4
WSD	Loureiro and Jorge (2019)	67.7
Ensemble	Gari Soler et al (2019)	66.7
BERT-large	WiC's paper	65.5
ELMo-weighted	Ansell et al (2019)	61.2
Context2vec	WiC's paper	59.3
Elmo	WiC's paper	57.7
Sense representations		
DeConf	WiC's paper	58.7
SW2V	WiC's paper	58.1
JBT	WiC's paper	53.6
Sentence level baselines		
Sentence Bag-of-words	WiC's paper	58.7
Sentence LSTM	WiC's paper	53.1
Random baseline		50.0



WiC Leaderboard

Human performance:
80% (Acc)



Contextualised word embeddings	Implementation	Accuracy %
BERT-large	Wang et al (2019)	68.4
WSD	Loureiro and Jorge (2019)	67.7
Ensemble	Gari Soler et al (2019)	66.7
BERT-large	WiC's paper	65.5
ELMo-weighted	Ansell et al (2019)	61.2
Context2vec	WiC's paper	59.3
Elmo	WiC's paper	57.7
Sense representations		
DeConf	WiC's paper	58.7
SW2V	WiC's paper	58.1
JBT	WiC's paper	53.6
Sentence level baselines		
Sentence Bag-of-words	WiC's paper	58.7
Sentence LSTM	WiC's paper	53.1
Random baseline		50.0



WiC in SuperGLUE

SuperGLUE is a benchmark consisting of a set of **challenging language understanding tasks**.

It includes tasks tackling different language phenomena such as **reading comprehension, question answering, language inference or co-reference**.

WiC is the task which measures the capacity of models for modeling **ambiguity**.

Room for improvement in SuperGLUE! **BERT score 69.0 vs 89.8 Human baseline**



But, two days ago...











But, two days ago in SuperGLUE leaderboard...

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines	↗	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
2	Facebook AI	RoBERTa	↗	84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	91.0/78.1	57.9
3	SuperGLUE Baselines	BERT++	↗	71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	99.4/51.4	38.0
		BERT	↗	69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	97.8/51.7	23.0
		Most Frequent Class	↗	47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	100.0/50.0	0.0
		CBoW	↗	44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	100.0/50.0	-0.4
		Outside Best	↗	-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]	↗	-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	-	47.6



RoBERTa (Facebook AI)

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
2	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	91.0/78.1	57.9
3	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	99.4/51.4	38.0
		BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	97.8/51.7	23.0
		Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	100.0/50.0	0.0
		CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	100.0/50.0	-0.4
		Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	-	47.6

RoBERTa has outperformed BERT by over 13 overall points in SuperGLUE!





RoBERTa (Facebook AI)

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines	🔗	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
2	Facebook AI	RoBERTa	🔗	84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	91.0/78.1	57.9
3	SuperGLUE Baselines	BERT++	🔗	71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	99.4/51.4	38.0
		BERT	🔗	69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	97.8/51.7	23.0
		Most Frequent Class	🔗	47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	100.0/50.0	0.0
		CBoW	🔗	44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	100.0/50.0	-0.4
		Outside Best	🔗	-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]	🔗	-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	-	47.6

However, not so large improvement in WiC (still under 70% and far from human performance)



Thanks!

Download training data from:

- <http://pilehvar.github.io/wic/>

Evaluate your model at:

- <https://competitions.codalab.org/competitions/20010>

WiC is part of  **SuperGLUE** benchmark!

- <https://super.gluebenchmark.com>