WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context

Anna Breit[1], Artem Revenko[1], Kiamehr Rezaee[2], Mohammad Taher Pilehvar[3], Jose Camacho-Collados[4]

1 2 3 4

https://github.com/semantic-web-company/wic-tsv

Word Sense Disambiguation

Find the most suitable sense:

S: (n) **mouse**
- S: (n) rodent, gnawer ✗
- S: (n) **mouse**, computer mouse ✓
  - S: (n) electronic device
- S: (n) shiner, black eye, **mouse** ✗
  - S: (n) bruise, contusion
- S: (n) **mouse** ✗
  - S: (n) person, individual

*"a **mouse** takes much more room than a trackball"*

Can we have ALL the senses?

*"I had a very interesting interview with Terno Schwab, the CEO of **Alphabet**."*

If we would have only one sense, it will always be the most suitable sense...

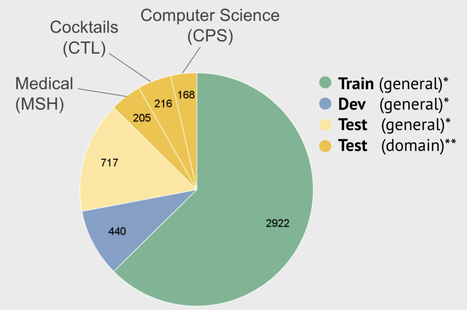
Target Sense Verification

Verify against ONE target sense:

Alphabet ? *"I had a very interesting interview with Terno Schwab, the CEO of **Alphabet**."*

+ no need to model the entire sense inventory
+ no assumption of complete data
→ more suitable for enterprise and domain-specific use cases

WiC-TSV Benchmark



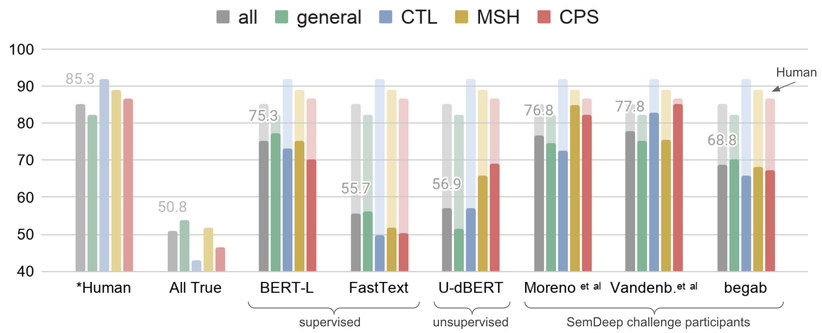
WiC-TSV allows to:
- ✓* investigate generalisation capabilities
- ✓**investigate the ability to transfer intrinsic knowledge (from general domain) into domain specific settings

Examples

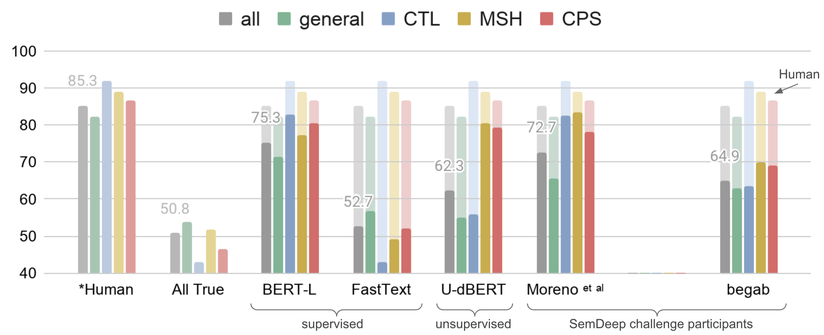
Tag	Context	Definition	Hypernyms
General Purpose (WNT/WKT)			
T	**Smoking** is permitted .	the act of smoking tobacco or other substances	breathing; respiration; ventilation
Cocktails (CTL)			
F	After a morning 's work I went off to see the **Bellini** retrospective at the Quirinale . Beautiful !	A Bellini cocktail is a mixture of Prosecco sparkling wine and peach purée.	cocktail
Medical Subjects (MSH)			
F	**Corona** Labs is happy to announce the general availability of the public beta of Android 64-bit Corona builds .	A viral disorder characterized by high fever; cough; dyspnea; renal dysfunction and other symptoms of a viral pneumonia.	viral_ pneumonia; coronavirus_ infection
Computer Science (CPS)			
T	pandas is a fast , powerful and easy to use open source data analysis and manipulation tool , built on top of the **Python** programming language .	Python is an interpreted, high-level, general-purpose programming language	object_ oriented_ programming_ language

Results

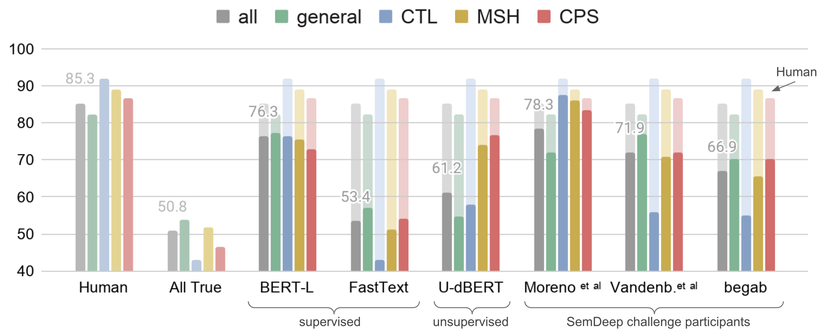
Task 1: Only Definition provided (Accuracy vs Human Accuracy)



Task 2: Only Hypernyms provided (Accuracy vs Human Accuracy)



Task 3: Hypernyms and Definition provided (Accuracy vs Human Accuracy)



Few-shot in-domain Analysis (Accuracy)

	Cocktails		Medical		Computer Science	
Trained on:	*seen*	unseen	*seen*	unseen	*seen*	unseen
General domain	*74.7*	75.5	82.9	76.0	77.8	74.8
General domain + Few shot in Domain	*+14.8*	+0.0	*+3.3*	+0.6	*+11.1*	-1.9
Few shot in Domain	*+14.1*	+5.1	*+5.7*	-0.5	*+15.5*	-5.7

Accuracy of BERT-L model on Task 3 when (1) continuing fine-tuning (2) only fine-tuning on 100 domain-specific instances, for target senses seen / unseen during training

This work was partially funded by Prêt-à-LLOD (https://pret-a-llod.github.io/), an EU Horizon 2020 project funded under grant agreement No. 825182.