XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization

https://pilehvar.github.io/xlwic/

Alessandro Raganato* Tommaso Pasini* Jose Camacho-Collados Mohammad Taher Pilehvar

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

SAPIENZA NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Tehran Institute for Advanced Studies

EMNLP 2020

WiC: The Word-in-Context Task

- **Binary classification task:** recognising whether a target word is used with the same meaning or not in two different contexts.

WiC: The Word-in-Context Task

- **Binary classification task:** recognising whether a target word is used with the same meaning or not in two different contexts.

Target word: **bed.**

Context-1: There’s a lot of trash on the **bed** of the river.

Context-2: I keep a glass of water next to my **bed** when i sleep.

XL-WiC: The Multilingual Benchmark

So far WiC benchmark (featured as a part of the SuperGLUE benchmark, and for a shared task at SemDeep-5 IJCAI workshop) - English only.

XL-WiC extends the WiC dataset to **12 new languages** from different families and with different degrees of resource availability:

- Bulgarian (BG)
- Danish (DA)
- German (DE)
- Estonian (ET)
- Farsi (FA)
- French (FR)
- Croatian (HR)
- Italian (IT)
- Japanese (JA)
- Korean (KO)
- Dutch (NL)
- Chinese (ZH)

XL-WiC: The Multilingual Benchmark

So far WiC benchmark (featured as a part of the SuperGLUE benchmark, and for a shared task at SemDeep-5 IJCAI workshop) - English only

XL-WiC extends the WiC dataset to **12 new languages** from different families and with different degrees of resource availability:

Lang	Target	Context-1	Context-2	Label
EN	Beat	We beat the competition	Agassi beat Becker in the tennis championship.	True
DA	Tro	Jeg tror p˚a det, min mor fortalte.	Maria troede ikke sine egne øjne.	True
ET	Ruum	Uhel hetkel olin v̈ aljaspool aega ja ruumi.	Umberringi oli Ĩ oputu ẗ uhi ruum.	True
FR	Causticité	Sa causticité lui a fait bien des ennemis.	La causticité des acides.	False
KO	틀림	틀림이 있는지 없는지 세어 보시오.	그 아이 하는 짓에 틀림이 있다면 모두 이 어미 죄이지요.	False
ZH	發	建築師希望發大火燒掉城市的三分之一。	如果南美洲氣壓偏低，則印度可能發乾旱	True
FA	صرف	صرف غذا نیم ساعت طول کشید	معلم صرف افعال ماضی عربی را آموزش داد	False

XL-WiC: The Multilingual Benchmark

XL-WiC was built using example sentences from two type of resources:
WordNet and **Wiktionary**

Global WordNet Association

Wiktionary
The free dictionary

XL-WiC: The Multilingual Benchmark - WordNet

Global WordNet Association

- Example usages from 9 languages:
    - Bulgarian, Chinese, Croatian, Danish, Dutch, Estonian, Japanese, Korean and Farsi
- We provide **dev** and **test** data for each language

XL-WiC: The Multilingual Benchmark - WordNet

- Example usages from 9 languages:
  - Bulgarian, Chinese, Croatian, Danish, Dutch, Estonian, Japanese, Korean and Farsi.
- We provide **dev** and **test** data for each language.

★ WordNet is known to be a fine-grained resource:
  - often different senses of the same word are hardly distinguishable from one another even for humans.

XL-WiC: The Multilingual Benchmark - WordNet

Global WordNet Association

- ★ WordNet is known to be a fine-grained resource:
  - ○ often different senses of the same word are hardly distinguishable from one another even for humans.

➔ Filtering step:

- ◆ automatic pruning of subtle sense distinctions leveraging WordNet structure itself. We removed all pairs whose senses were:
  - ✂ first degree connections in the WordNet semantic graph;
  - ✂ sister senses;
  - ✂ belonged to the same supersense.

XL-WiC: The Multilingual Benchmark - WordNet

- Case study: Farsi
  - make a challenging dataset with sense distinctions that are easily interpretable by humans.

➔ Semi-automatic extraction:
  - we extracted all example usages and asked an annotator to group them into positive and negative pairs.

XL-WiC: The Multilingual Benchmark - Wiktionary

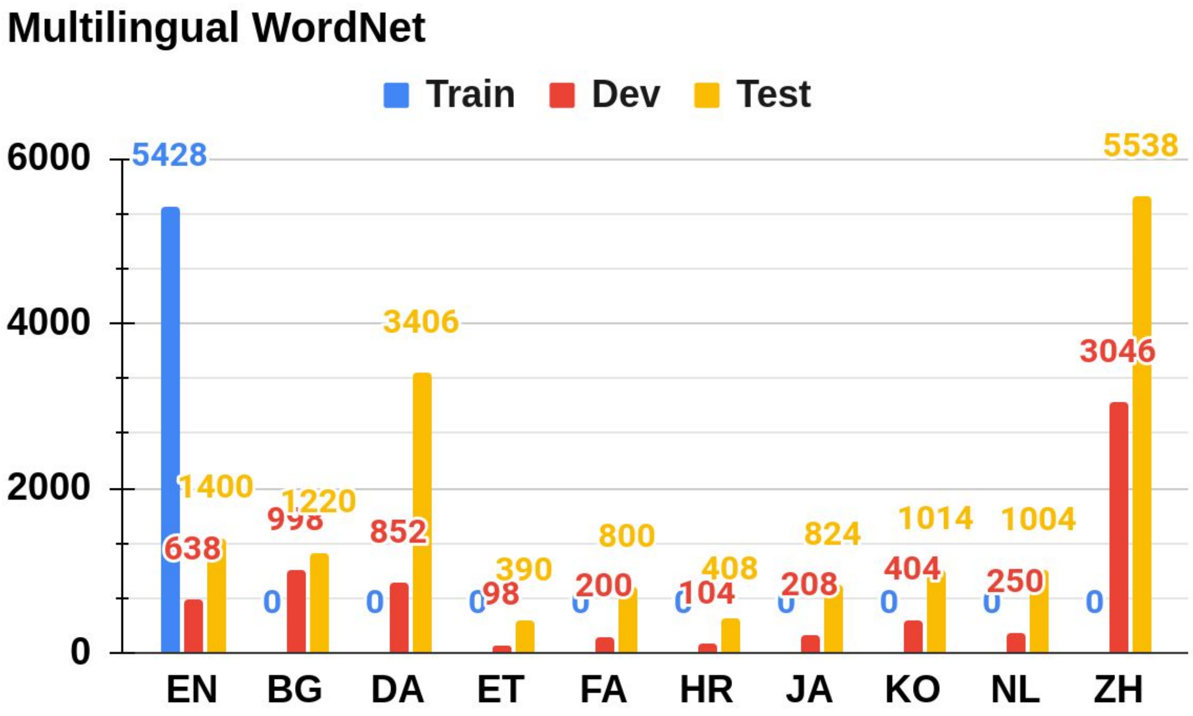
XL-WiC: The Multilingual Benchmark - Wiktionary

Wiktionary
The free dictionary

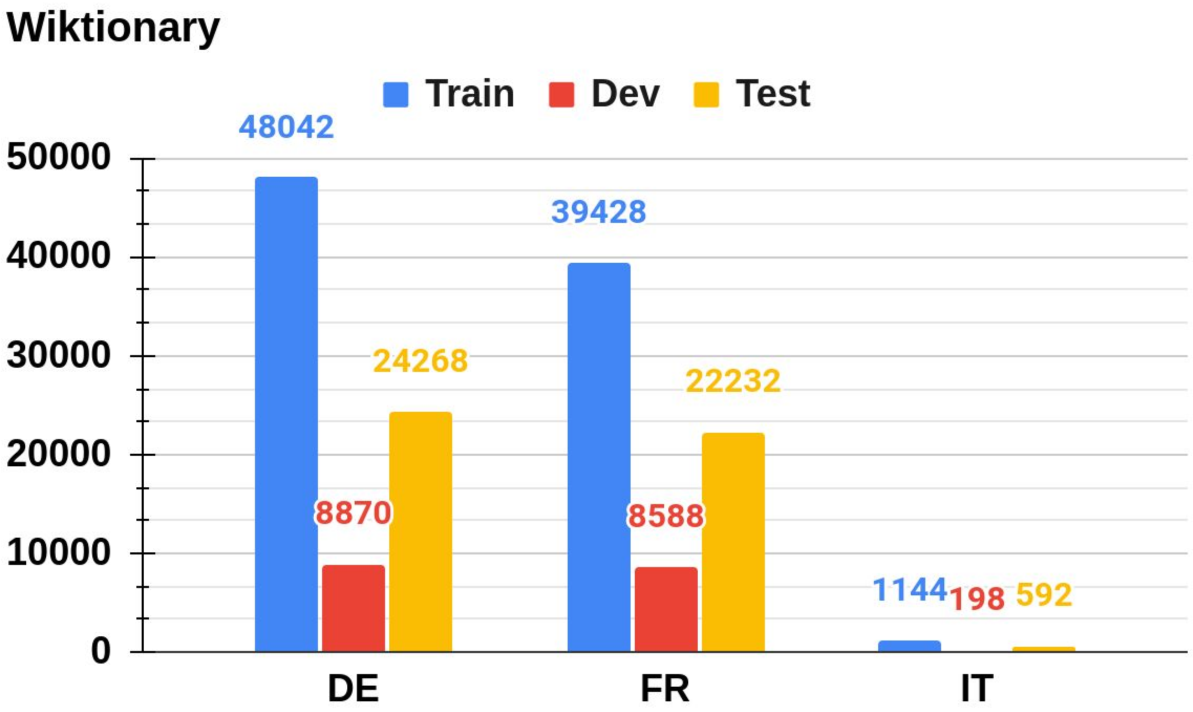
- We extracted examples for three European languages for which we did not have WordNet-based data:
  - French, German, and Italian.

- We provide **train**, **dev** and **test** data for each language.

XL-WiC: The Multilingual Benchmark - Statistics



XL-WiC: The Multilingual Benchmark - Statistics



XL-WiC: human performance

WiC	WordNet						Wiktionary	
EN	DA	FA	IT	JA	KO	ZH	DE	IT
80.0*	87.0	97.0	82.0	75.0	76.0	85.0	74.0	78.0

Human performance (in terms of accuracy) on 100 random instances for different languages in XL-WiC .

XL-WiC: Experimental setup

- Models:
  - mBERT
  - XLM-R-base
  - XLM-R-large
  - L-BERT (language specific BERT models used in monolingual setting)

- Evaluation settings:
  - Cross-Lingual Zero-shot
  - Multilingual Fine-Tuning
  - Monolingual
  - Translation

- Evaluation metric:
  - accuracy

XL-WiC: Experimental setup

- Models:
  - mBERT
  - XLM-R-base
  - XLM-R-large
  - L-BERT (language specific BERT models used in monolingual setting)

- Evaluation settings:
  - **Cross-Lingual Zero-shot**
  - Multilingual Fine-Tuning
  - **Monolingual**
  - Translation

- Evaluation metric:
  - accuracy

XL-WiC: Experiments and results

- Cross-Lingual Zero-shot: WordNet

Model	**BG**	**DA**	**ET**	**FA**	**HR**	**JA**	**KO**	**NL**	**ZH**
Zero-shot cross-lingual setting ***Train***: *EN* – ***Dev***: *EN*									
mBERT	58.28	64.86	62.56	71.50	63.97	62.26	59.76	63.84	69.36
XLMR-base	60.73	64.79	62.82	69.88	62.01	60.44	66.96	65.73	65.78
XLMR-large	**66.48**	**71.11**	**68.71**	**75.25**	**72.30**	63.83	**69.63**	**72.81**	**73.15**

XL-WiC: Experiments and results

- Cross-Lingual Zero-shot: WordNet

Model	BG	DA	ET	FA	HR	JA	KO	NL	ZH
				Zero-shot cross-lingual setting			***Train**: EN*	–	***Dev**: EN*
mBERT	58.28	64.86	62.56	71.50	63.97	62.26	59.76	63.84	69.36
XLMR-base	60.73	64.79	62.82	69.88	62.01	60.44	66.96	65.73	65.78
XLMR-large	**66.48**	**71.11**	**68.71**	**75.25**	**72.30**	63.83	**69.63**	**72.81**	**73.15**

- Cross-Lingual Zero-shot: Wiktionary

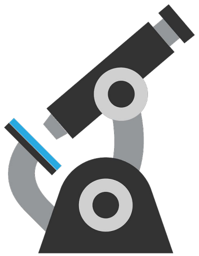
	Model	DE	FR	IT
Z-Shot	mBERT	58.27	56.00	58.61
	XLMR-base	58.30	56.13	55.91
	XLMR-large	**65.83**	**62.50**	**64.86**

XL-WiC: Experiments and results

- Monolingual: Wiktionary

	Model	DE	FR	IT
Mono	mBERT	81.58	73.67	71.96
	XLMR-base	80.84	73.06	68.58
	XLMR-large	**84.03**	76.16	72.30
	L-BERT	82.90	**78.14**	**72.64**

XL-WiC: Analysis

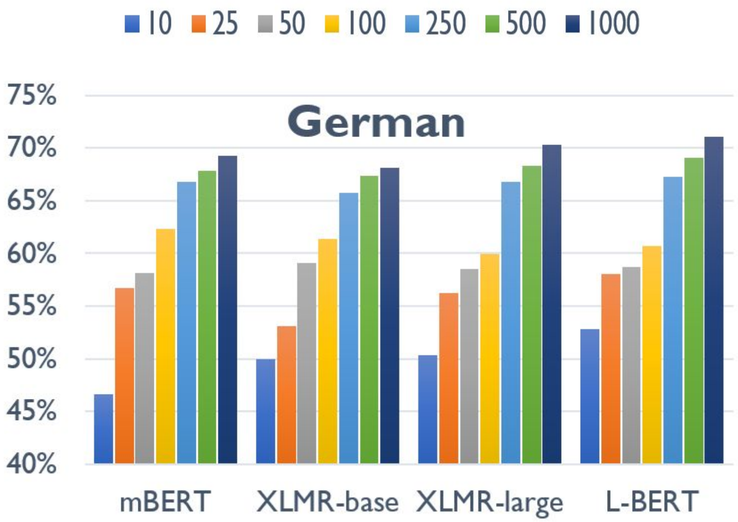


- Seen (IV) and Unseen (OOV) Words:
  - IV: is a test set containing instances where the target words **were seen** at **training time**.
  - OOV: is a test set containing instances where the target words **were NOT seen** at **training time**.

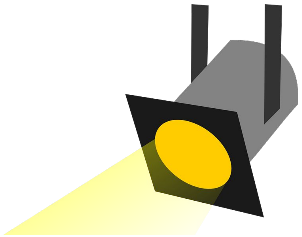
	Model	DE	FR	IT
IV	mBERT	81.86	72.92	73.15
	XLMR-base	81.17	71.92	70.69
	XLMR-large	**84.24**	75.61	**75.12**
	L-BERT	83.23	**77.62**	73.89
OOV	mBERT	70.08	71.24	68.54
	XLMR-base	71.31	71.14	62.36
	XLMR-large	72.54	73.93	65.17
	L-BERT	**76.64**	**78.00**	**69.10**

XL-WiC: Analysis

- Few-shot Monolingual:
  - Evaluation performed when training each model on 10, 25, 50, 100, 250, 500, 1000 annotated examples.



XL-WiC: Conclusions



- **XL-WiC**: a large benchmark with **over 80K instances** for evaluating context-sensitive models in **13 languages**!

- Testbed for **cross-lingual experimentation** in settings such as zero-shot or few-shot transfer across languages.

- We provide **performance baselines with current multilingual neural language models**, showing that **there is still room for improvement**, especially for languages such as Japanese, Korean or Farsi in the cross-lingual zero-shot setting.

https://pilehvar.github.io/xlwic/

https://pilehvar.github.io/xlwic/

Thank you!