Multilingual and Cross-lingual Word-in-Context Disambiguation

Federico Martelli, Najla Kalach, Gabriele Tola and Roberto Navigli
Sapienza University of Rome

MOUSSE

Main contribution

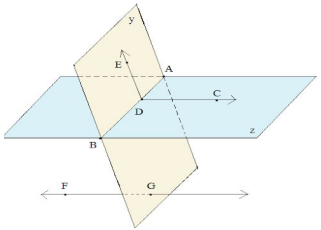
- A novel entirely manually-annotated **multilingual and cross-lingual evaluation benchmark**
- **Goal:** testing the **systems'** ability to **discriminate contextual meanings without** relying on a **fixed sense inventory**
- 5 languages: English, French, Arabic, Russian and Chinese.

Word Sense Disambiguation (WSD)

Identifying the meaning of a word in given context

the **plane** takes off

?

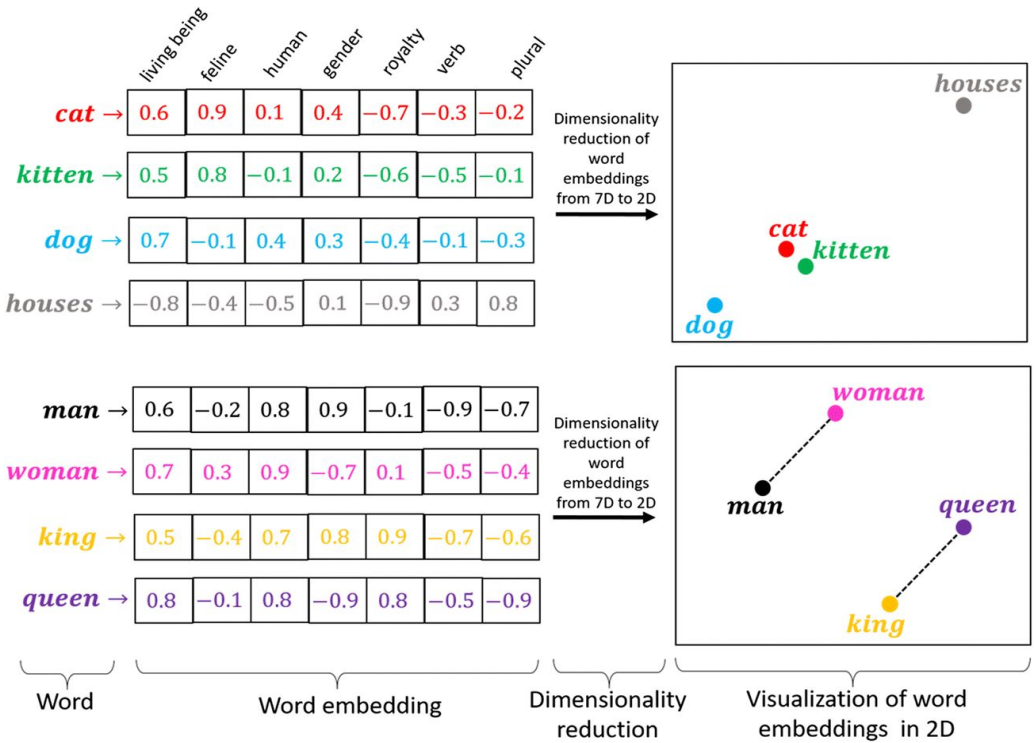


Limits of WSD

- **Limited availability** of wide coverage sense inventories (mainly WordNet)
- **Fine granularity** of the sense inventory
  - S: (v) **run** (move fast by using one's feet, with one foot off the ground at any given time) *"Don't run--you'll be out of breath"; "The children ran to the store"*
  - S: (v) **run** (travel rapidly, by any (unspecified) means) *"Run to the store!"; "She always runs to Italy, because she has a lover there"*

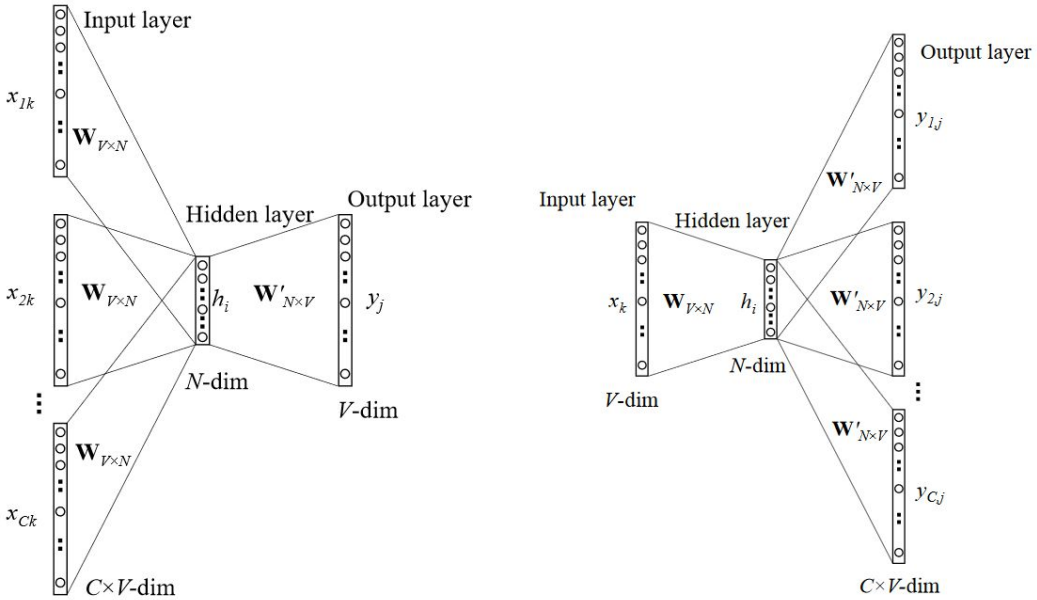
A possible solution is to do away with explicit word senses...

Once only humans could understand meaning…



4 - SemEval-2021 - Multilingual and Cross-lingual Word-in-Context Disambiguation
Martelli, Kalach, Tola, Navigli

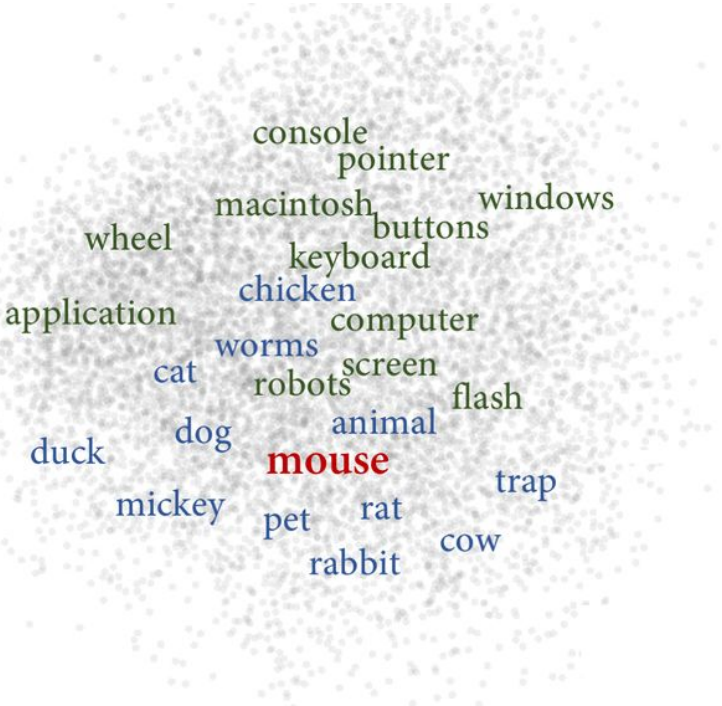
Word embeddings



- encode low-dimensional vectors from **corpora**,
- proved to have **generalization power**, and
- encode **different meanings into the same representation**.

Meaning conflation deficiency

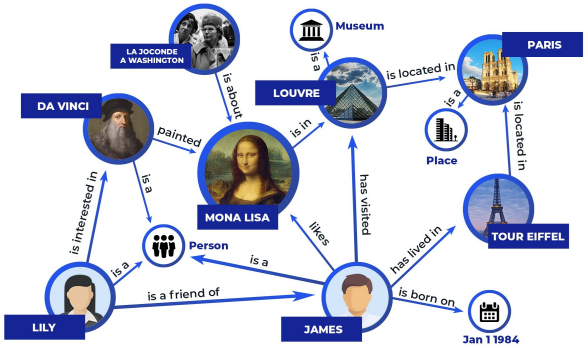
- **Incapability of distinguishing** between different **meanings** of a word, and
- Consider the word **mouse**: the animal meaning and the device meaning are merged into one single representation.



Meaning representations

UNSUPERVISED APPROACHES

KNOWLEDGE-BASED APPROACHES

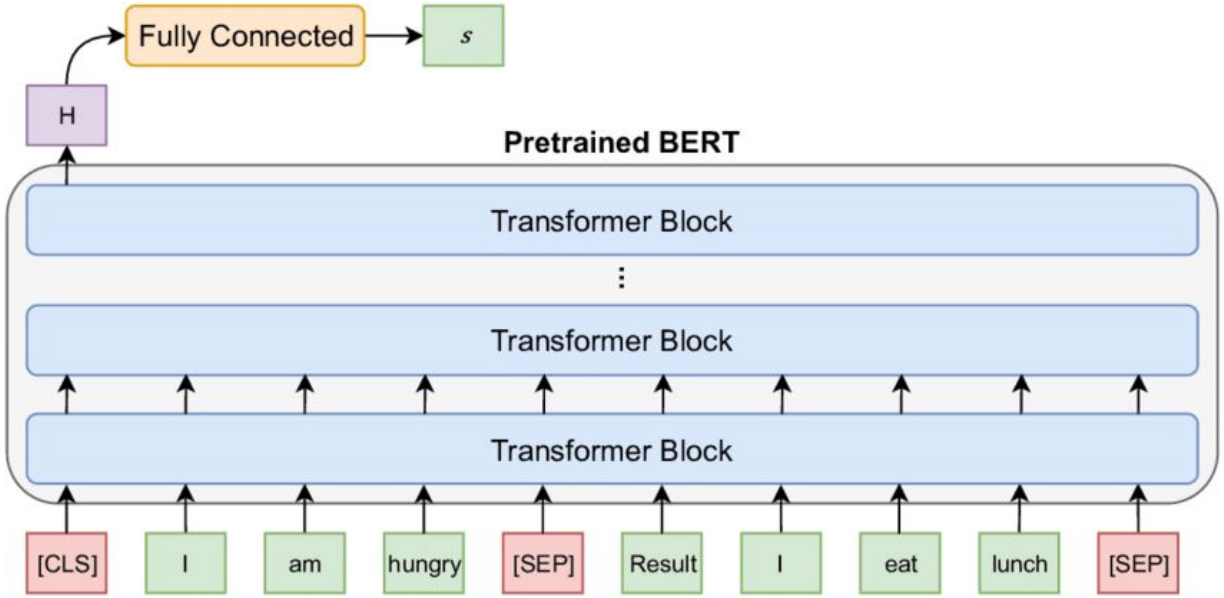


7 - SemEval-2021 - Multilingual and Cross-lingual Word-in-Context Disambiguation
Martelli, Kalach, Tola, Navigli

SAPIENZA

Contextualized embeddings

- are **sensitive to the context**, and
- allow to take account **of the graded effects in context**.



But how to provide an alternative evaluation?

- Testing explicit and implicit WSD **intrinsically**
- **Dropping** the requirement of a **fixed sense inventory**
- Enabling **multilingual** and **cross-lingual evaluation**

Durch das Schätzen erst giebt es Werth: und ohne das Schätzen wäre die Nuss des Daseins hohl.

"It is only through evaluation that value exists: And without evaluation the nut of existence would be hollow."

Also Sprach Zarathustra
Friedrich Nietzsche

State of the art

Stanford Contextual Word Similarity (SCWS)
Huang et al., ACL 2012

Word-in-Context (WIC)
Pilehvar and Collados, NAACL-HLT 2019

CoSimLex
Armendariz et al., LREC 2020

The WiC dataset at a glance

The WiC dataset frames the evaluation as a **binary classification task**.

Given two sentences, in which a target word *W* occurs, a system is asked to provide a binary answer (**T/F:** same meaning or not).

F	**play**	V	0-0	**Play** fair .	**Play** football .

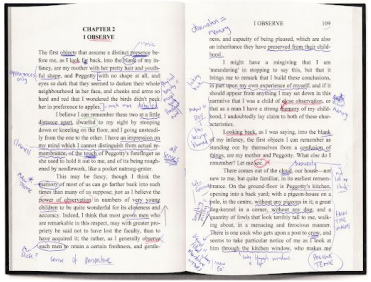
The WiC dataset at a glance

- enables intrinsic evaluation
- straightforward take on WSD
- only nouns and verbs
- no manual annotation
- no multilingual and cross-lingual data

Dataset composition

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

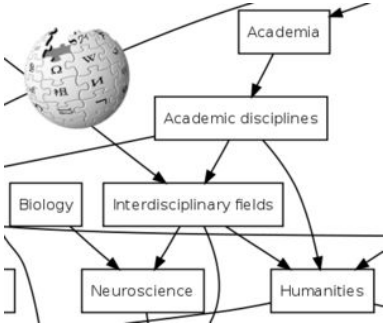
Desiderata



Multilinguality & Cross-linguality

Manually-annotated data

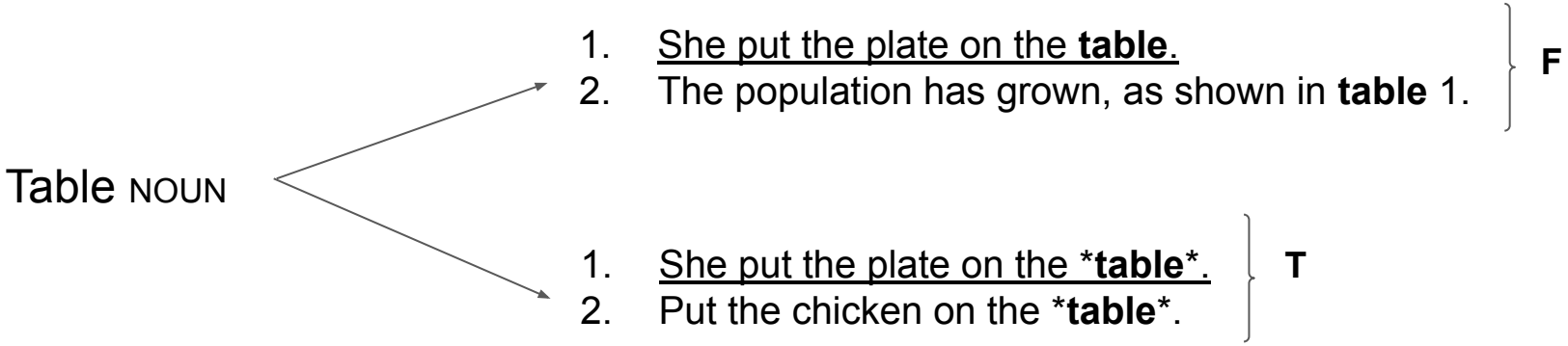
Genre & domain and POS coverage



Multilingual and Cross-lingual Word-in-Context (MCL-WiC)

MCL-WiC in a nutshell

- first SemEval task for Word-in-Context disambiguation
- structured as **binary classification task**,
- enables **multilingual & cross-lingual evaluation**,
- cover **all parts of speech** as well as **different genres and domains,**
- for each target lemma-pos: two sentence pairs (<u>one sentence fixed</u>), and
- **50%** of the data is derived from **UNPC**, **50%** of the data from **Wikipedia**



MCL-WiC in a nutshell

The MCL-WiC dataset is composed of two parts:

MULTILINGUAL
Pairs of sentences in the same language

CROSS-LINGUAL
Pairs of sentences in different languages

MCL-WiC: Multilingual part

1. In that context of coordination and integration, Bolivia holds a key *play* in any process of infrastructure development.
2. A musical *play* on the same subject was also staged in Kathmandu for three days.

1. Во время обсуждения любого вопроса каждый представитель может внести предложение о перерыве или закрытии *заседания*.
2. Он сомневается в целесообразности такого решения и считает, что планирование заседаний Пятого комитета не должно зависеть от работы пленарных *заседаний*.

MCL-WiC: Cross-lingual part

1. Any alterations which it is proposed to make as a result of this review are to be *reported* to the Interdepartmental Committee on Charter Repertory for its approval.
2. Les participants *feront rapport* directement aux autorités compétentes de leurs pays.

1. Any alterations which it is proposed to make as a result of this review are to be *reported* to the Interdepartmental Committee on Charter Repertory for its approval.
2. Участники *представят отчеты* соответствующим органам в своих странах.

United Nations Parallel Corpus (UNPC)

- Official documents manually **translated**
- **6 languages** (English, Chinese, French, Spanish, Russian and Arabic)
- **11 millions sentences** per language
- **86,000 documents**

Wikipedia

- a multilingual open-collaborative online encyclopedia
- created and maintained by a community of volunteer editors

Annotation process: workbench

Lemma: parade.NOUN	
1	The author further alleges that, before he was put on the identification *__parade__*, he was taken to his house to shower, shave and dress, as instructed by the police.
F	This oath is reiterated at every formal meeting, *__parade__* or military gathering.
F	Both nationalists and unionists hold *__parades__* and marches, but the vast majority is sponsored by loyalist orders.

Annotation process: workbench

Linguistic experts are provided with:

- **a list of target lemmas** selected starting from a manually-created lexicon,
- **a list of potential sentences** (extracted from UNPC) for each target lemma, and
- **instructions** regarding the usage of the **Wikipedia search engine** for the selection of sentences from Wikipedia.

Annotation process: criteria

Linguistic experts are asked to:

- annotate with the tags **T** (True) or **F** (False),
- use the tag **T** only if the **meaning** of the target words occurring in the two sentences are **exactly the same**,
- use the tag **F in all the remaining cases** (even in the case of figurative usages),
- **use reputable dictionaries** to discriminate senses and, most important avoid subjectivity, and
- discard not well-formatted and semantically unclear sentences.

Overview: multilingual

Number of sentences

Language	Training	Development	Test
English	16,000	2,000	2,000
Chinese		2,000	2,000
French		2,000	2,000
Russian		2,000	2,000
Arabic		2,000	2,000

Overview: multilingual

Number of unique lemmas

Language	Training	Development	Test
English	4,000	500	500
Chinese		500	500
French		500	500
Russian		500	500
Arabic		500	500

Overview: cross-lingual

Number of sentences

Lang. comb.	Training	Development	Test
EN-AR			2,000
EN-FR			2,000
EN-RU			2,000
EN-ZH			2,000

Overview: cross-lingual

Number of unique lemmas

Lang. comb.	Training	Development	Test
EN-AR			500
EN-FR			500
EN-RU			500
EN-ZH			500

Baselines

1) **sense embeddings**, such as LMMS (Loureiro and Jorge, 2019) and SensEmBERT (Scarlini et al., 2020), which combine contextualized embeddings with the knowledge derived from resources such as WordNet and BabelNet;

2) **context-specific word embeddings**, such as Context2vec (Melamud et al., 2016), BERT (Devlin et al., 2019) etc.

Linguistic issues

Linguistic issues - Arabic

Diacritics: same pronunciation vs same meaning

a) Same writing, but different pronunciation and meanings

دين dīn (religion)

دين dayn (debt)

b) Same writing and pronunciation, but different meanings

ظرف ẓarf (adverb)

ظرف ẓarf (envelope)

Linguistic issues - Arabic

Some non-vocalized Arabic words have several possible

Unvocalized target word

ذهبت البنت إلى المدرسة.

The girl went to school.

كانت حقيبة المدرسة جديدة.

The teacher's bag was new.

Vocalized target word

ذَهَبَت البِنْتُ إلى المَدْرَسَةِ.

Dhabat al-bint^{u} 'ila al-madrasat^{i}.

The girl went to school

كانَت حقيبةُ المُدَرِّسَةِ جَدِيدَةً.

Kānat ḥaqībat^{u} al-mudarrisat^{i} jadīdat^{an}.

The teacher's bag was new.

Linguistic issues - Chinese

1) All homophones are basically lost. If two unrelated words are pronounced in the same way, such as “plane” (the airplane) and “plane” (the surface), they will never be written in the same way.

Chénmò: 沉默 (“silent; to be silent”) and 沉没 (“to sink”)

Linguistic issues - Chinese

2) Some ambiguity is lost in translation. E.g., in Chinese there is the category of classifiers (neither nouns nor articles) and that has many potentialities for ambiguity. But they are basically not found in “European” languages and therefore are not appliable in a comparative task:

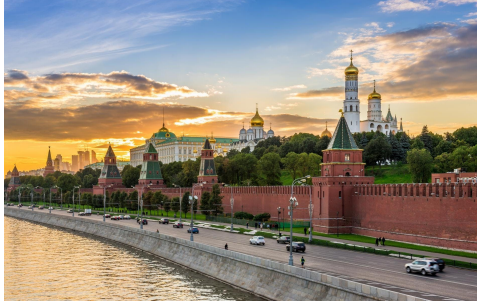
道dào, 1) classifier for long and narrow object; 一道河 a river (one+classifier+river)and other meanings as well

Linguistic issues - Chinese

3) Some “words” may appear ambiguous, but in practice they are not.

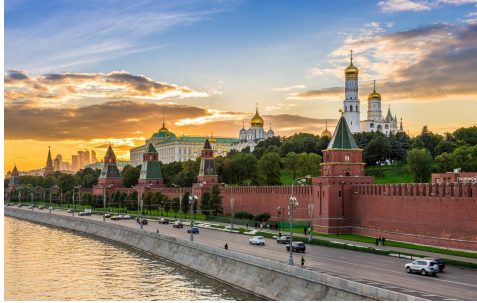
果 guǒ appears with at least two meaning in dictionaries: “fruit” and “result”, but this character is never a word itself, since most of the Chinese words are in the contemporary lexicon composed of two or more characters; when it appears in actual text, guǒ is connected to another character, and the word thus formed is not ambiguous anymore.

Linguistic issues - Russian



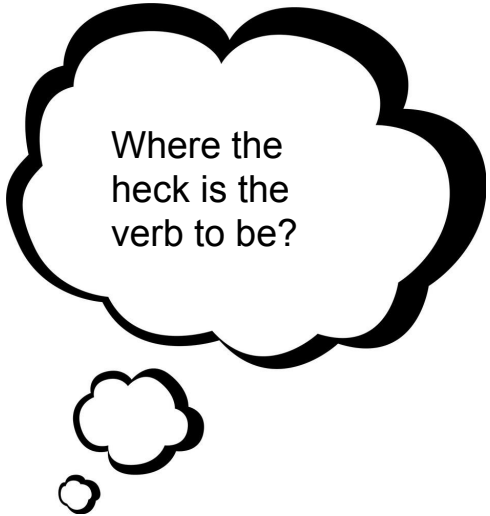
Lemma: be.VERB	
1	He always wanted to **be** a teacher.
?	Она преподаватель. (She is a teacher)

Linguistic issues - Russian

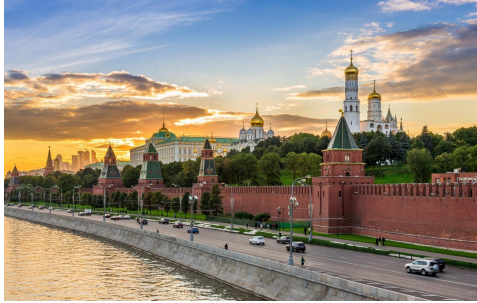


Она преподаватель

She teacher

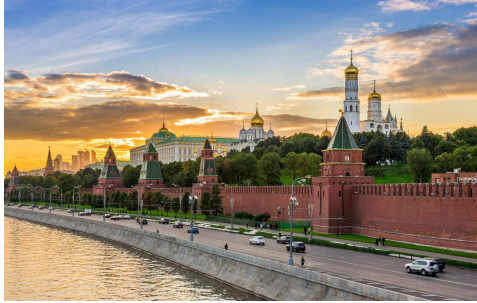


Linguistic issues - Russian



Lemma: have.VERB	
1	I **have** a new house.
?	У меня красивая машина. (I’ve got a beautiful car)

Linguistic issues - Russian



У меня красивая машина.

By me beautiful car

?

Data format - 1


```
[
   {
                "id": "dev.en-en.0",
                "lemma": "play",
                "pos": "NOUN",
                "sentence1": "In that context of coordination and integration, Bolivia holds a key play
in any process of infrastructure development.",
                "sentence2": "A musical play on the same subject was also staged in Kathmandu for
three days.",
                "start1": "69",
                "end1": "73",
                "start2": "10",
                "end2": "14",
   }
]
```

Data format - 2


```
[
        {
                "id": "dev.en-en.0",
                "tag": "F"
        }
]
```

To sum up

We propose:

- a novel multilingual and cross-lingual dataset in **5 languages which**
- enables **inherent evaluation** of approaches for WSD,
- in **multilingual** and **cross-lingual** settings, and
- fully **manually annotated**.

..wanna participate?!

REGISTER TO: CodaLab

PARTICIPATE IN: **https://competitions.codalab.org/competitions/27054**

AND:

References

- Pilehvar, Mohammad Taher, and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. Proceedings of NAACL-HLT 2019, pages 1267-1273.
- Loureiro, Daniel and Alipio, Jorge. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pages 5682-5691.
- Scarlini, Bianca; Pasini, Tommaso and Navigli, Roberto. SensEmBERT: Context-Enhanced SenseEmbeddings for Multilingual Word Sense Disambiguation, Proceedings of the Association for the Advancement of Artificial Intelligence, 2020, pages 8758-8765.
- Melamud, Oren; Goldberger, Jacob and Dagan, Ido. context2vec:Learning Generic Context Embeddingwith Bidirectional LSTM Proceedings of the 20th SIGNLL conference on computational natural language learning, 2016, pages 51-61.
- Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton and Toutanova, Kristina BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Proceedings of NAACL-HTL 2019, pages 4171-4186.
- Camacho-Collados, José; Pilehvar, Mohammad Taher, From word to sense embeddings: a survey on vector representations of meaning, Journal of Artificial Intelligence Research, 63, 2018, pages 743-788.

https://competitions.codalab.org/competitions/27054