WiC-ITA@EVALITA 2023: Task Guidelines

Pierluigi Cassotti
Dept. of Computer Science
University of Bari, Italy
pierluigi.cassotti@uniba.it

Lucia Siciliani
Dept. of Computer Science
University of Bari, Italy
lucia.siciliani@uniba.it

Lucia Passaro
Dept. of Computer Science
University of Pisa, Italy
lucia.passaro@unipi.it

Maristella Gatto
Dip. di Ricerca e Innovazione Umanistica
University of Bari, Italy
maristella.gatto@uniba.it

Pierpaolo Basile
Dept. of Computer Science
University of Bari, Italy
pierpaolo.basile@uniba.it

1 Background

Word Sense Disambiguation (WSD) (Bevilacqua et al., 2021) is a Natural Language Processing task with a long history and of extremely interesting for the Computational Linguistics community. In WSD the goal is to disambiguate each word occurrence assigning them the correct sense from a fixed sense inventory, such as WordNet (Miller, 1992). The introduction of contextualized models, such as BERT, allowing the representation of a word in different contexts, steers the research focus to new tasks, such as the Word in Context (WiC) task (Pilehvar and Camacho-Collados, 2019).

WSD and the WiC task are highly related: while the former models in an explicit way the relationship between the target word and its sense (taken from a pre-defined sense inventory), the latter reduces it to a binary task. The WiC task requires determining if a word occurring in two different sentences has the same meaning or not. In recent years, there has been a growing interest in the WiC task, demonstrated by the creation of several different resources and shared tasks covering more than 20 languages, shown in Table 1.

Dataset	**Languages**
WiC	EN
XL-WiC	EN, BG, ZH, HR, DA, NL, ET, FA, FR, DE, IT, JA, KO
MCL-WiC	EN, AR, FR, RU, ZH EN-AR, EN-FR, EN-RU, EN-ZH
AM^2iCo	EN-DE, EN-RU, EN-JA, EN-KO, EN-ZH, EN-AR, EN-IN, EN-FI, EN-TR, EN-EU, EN-KA, EN-UR, EN-BN, EN-KK

Table 1: Languages distribution among datasets.

In general, the Word in Context task is of broad-scope interest, as it is not limited to specific domains and can be useful for several NLP tasks. Furthermore, the training and the evaluation on a monolingual (Italian) or cross-lingual (English-Italian) dataset is advantageous not only for the models for the Italian language. In fact, the transfer learning ability of WiC models across different languages is proven in previous works (Martelli et al., 2021), where models improve their performance by training in other languages. Several initiatives have been proposed throughout the years: the first one (Pilehvar and Camacho-Collados, 2019) being the proposal of the WiC task, which also came along with a dataset but was limited to English. For this reason, it was followed by the XL-WiC (Raganato et al., 2020) dataset which tried to tackle this issue by taking into account a total of 15 languages. Next, the MCL-WiC (Martelli et al., 2021) was the first WiC dataset to introduce the Cross-lingual task. The main motivation behind this particular choice was to cover scenarios where systems have to deal with different languages

simultaneously, further highlighting the importance of this task in real-world applications. With AM^{2}iCo (Liu et al., 2021) the main aim was to focus on low-resource languages and to ensure participating models must consider both the target word and the context to achieve good performances. Finally, in CoSimLex (Armendariz et al., 2020) the task is extended to *pairs* of words that appear in a shared context and the goal is to determine to which degree they refer to the same concept. This is done to capture the word polysemy as well as the context-dependency of words.

Shared tasks regarding the WiC usually preserve its binary design, where the two possible outcomes for each entry are: true if the meaning of the target word changes between the two sentences/contexts and false if it does not. However, there can be some cases where it is not so simple to determine the lack or presence of semantic similarity in a discrete way. For this reason, we exploit the 4-point relatedness scale introduced by (Schlechtweg et al., 2018; Brown, 2008) in the annotation process. The scale consists of 4 values, namely 4: Identical; 3: Closely Related; 2: Distantly Related; 1: Unrelated. A fifth value can be assigned (0: Cannot decide) for uncertain cases.

Unfortunately, as often happens in the Natural Language Processing research area, some languages are more represented than others, and the WiC task makes no exception in this sense. This issue is evident by analyzing Table 1, where only the XL-WiC dataset (Raganato et al., 2020) contains data for the Italian language. With the WiC-ITA task, we aim at filling this gap in the literature, making openly available a resource that can undoubtedly foster novel research.

2 Task Description

The general goal of the WiC-ITA task is to establish if a word w occurring in two different sentences s_1 and s_2 has the same meaning or not. In particular, our task is composed of two subtasks: the binary classification (Subtask 1) and the ranking (Subtask 2). Participants are allowed to participate in one or both of the subtasks.

2.1 Subtask 1: Binary Classification

Subtask 1 is structured as follows:
Given a word w occurring in two different sentences s_1 and s_2, the goal is to provide the sentences pair with a score determining whether w maintains the same meaning or not.
Possible outcomes for this subtask is:

- 0: the word w has *not* the same meaning in the two sentences s_1 and s_2;
- 1: the word w has the same meaning in the two sentences s_1 and s_2.

An example of output for subtask 1 is given in Listing 1.

2.2 Subtask 2: Ranking

Subtask 2 is structured as follows:
Given a word w occurring in two different sentences s_1 and s_2, the goal is to provide the sentences pair with a score indicating to which extent, in a 1-4 scale, w has the same meaning in the two sentences.
The scoring system for this subtask is a continuous value where $score \in [1, 4]$. An higher score corresponds to an higher degree of semantic similarity.

An example of output for subtask 2 is given in listing 2.

3 Development and Test Data

The creation of datasets for the WiC task usually relies on using sense inventories, such as WordNet or BabelNet (Navigli and Ponzetto, 2010). More specifically, sense inventories are exploited for selecting target words, which should exhibit polysemia and for the generation of sentences pairs using the sense examples provided, i.e. sentences in which the target word occurs with the respective sense. After the selection of target words and the generation of sentence pairs, only a small part of these are manually annotated/validated by human experts.

```
{
   "id": "ricevere.verb.1",
   "lemma": "ricevere",
   "sentence1": "( ANSA ) - BOLOGNA , 11 AGO - ' ' Parma ha ricevuto un
   altro grande riconoscimento e lo deve al Governo Berlusconi ed all'
   impegno dell' Ufficio scolastico regionale per l' Emilia - Romagna ' '
   .",
   "sentence2": "Ciascuno dei 1100 dipendenti della società
   riceverà un certo numero di azioni gratis , a seconda della
   posizione e del tempo trascorso nell' azienda .",
   "start1": 43,
   "end1": 51,
   "start2": 43,
   "end2": 51,
   "label" : 1
 }
 {
   "id": "cannone.noun.10",
   "lemma": "cannone",
   "sentence1": "Da li venivano dirette tutte le operazioni militari
   compresi i tiri di grossi cannoni posti in altri luoghi , ma sempre in
   alto .",
   "sentence2": "Il viaggio scorre che è un piacere , tra cannoni d'
   erba , caffè , colazioni varie , la tensione sembra non esistere .",
   "start1": 78,
   "end1": 85,
   "start2": 41,
   "end2": 48,
   "label": 0
 }
```


Listing 1: Subtask 1 Example.

Unlike previous datasets, for the WiC-ITA task, we do not rely on sense inventories, but we exploit unsupervised techniques for building both the set of target words and extracting the list of sentence pairs. Moreover, the human annotation will be carried out for *all* the sentence pairs, thus making WiC-ITA the largest manually annotated resource for the WiC task.

The WiC-ITA task presents two main differences from previous Word-in-Context tasks: (1) the WiC-ITA task does not rely on any sense inventory; (2) a massive annotation is carried out for the labeling of sentence pairs.

WiC-ITA includes monolingual (Italian) and cross-lingual (English-Italian) data. We provide data for the training, development, and test phases. In particular:

- the training and development set consists of annotated pairs of monolingual (Italian) sentences;
- the test set consists of annotated pairs of monolingual (Italian) sentences and annotated pairs of cross-lingual (English-Italian) sentences.

The monolingual and the cross-lingual pairs are extracted from the itWaC and ukWaC corpora, both part of the WaCKy project. ukWaC is a corpus obtained by crawling the web pages under the .uk domain. It consists of more than 2 billion words, annotated with PoS tags and lemmatized using the TreeTagger tool. itWaC, differently from ukWaC is lemmatized using Morph-it! and is obtained crawling web pages under the .it domain.

We associate with each sentence pair the average score assigned by the annotators according to the 4-point relatedness scale, the offsets of the target word on the respective sentences, and the lemma of the target word. We only consider the Italian lemma for the cross-lingual examples, but we provide the offsets for both languages. All the datasets are provided as JSON files. An example of the WiC-ITA dataset is sketched in Listings 2.

```
{
   "id": "ricevere.verb.1",
   "lemma": "ricevere",
   "sentence1": "( ANSA ) - BOLOGNA , 11 AGO - ' ' Parma ha ricevuto un
   altro grande riconoscimento e lo deve al Governo Berlusconi ed all'
   impegno dell' Ufficio scolastico regionale per l' Emilia - Romagna ' '
   .",
   "sentence2": "Ciascuno dei 1100 dipendenti della società
   riceverà un certo numero di azioni gratis , a seconda della
   posizione e del tempo trascorso nell' azienda .",
   "start1": 43,
   "end1": 51,
   "start2": 43,
   "end2": 51,
   "score" : 4.0
 }
 {
   "id": "cannone.noun.10",
   "lemma": "cannone",
   "sentence1": "Da li venivano dirette tutte le operazioni militari
   compresi i tiri di grossi cannoni posti in altri luoghi , ma sempre in
   alto .",
   "sentence2": "Il viaggio scorre che è un piacere , tra cannoni d'
   erba , caffè , colazioni varie , la tensione sembra non esistere .",
   "start1": 78,
   "end1": 85,
   "start2": 41,
   "end2": 48,
   "score": 1.0
 }
```

Listing 2: Subtask 2 Example.

3.1 Annotation

The gold truth of Subtask 2 consists of the average of the scores assigned by the annotators (from 4: Identical to 1: Unrelated).

- 4: the meaning of word w in the two sentences s_1 and s_2 is identical;
- 3: the meaning of word w in the two sentences s_1 and s_2 is closely related;
- 2: the meaning of word w in the two sentences s_1 and s_2 is distantly related;
- 1: the meaning of word w in the two sentences s_1 and s_2 is completely unrelated.

Each example is annotated by two independent annotators. **Annotations for which at least one of the annotators voted 0 (Cannot decide) are discarded from the annotated data.** The score for Subtask 2 is obtained by averaging the scores assigned by the two annotators.

The labels for the Sub Task 1 *(binary)* are obtained by exploiting only examples for which the two annotators agree. Two annotators agree in two cases:

1. both annotators give a score in the set $\{1, 2\}$;
2. both annotators give a score in the set $\{3, 4\}$;

In the first case, the example is labeled as 0, and in the second one as 1.

3.2 Limitations

The participants can exploit other data and resources for developing their systems. The only limitation is the usage of the two corpora adopted in the annotation of data: itWaC and ukWaC.

3.3 Data Release Policy

All the sentences are extracted by the itWaC and ukWaC corpora, part of the WaCKy project. The crawler used by the WaCKy project respects the download policies imposed by website administrators (i.e., the robots.txt file), and the WaCKy website contains information on how to request the removal of specific documents from the respective corpora. Moreover, we only provide samples of sentences from the corpus, making the reconstruction of the original document unfeasible.

4 Evaluation

We will provide rankings for each subtask and test set: Subtask 1 Monolingual; Subtask 1 Cross-lingual; Subtask 2 Monolingual; Subtask 2 Cross-lingual. We will now provide the details of the evaluation process for each subtask.

4.1 Subtask 1: Binary Classification

Systems' predictions will be evaluated against the gold truth using the F1-Score. In particular, the metrics used to evaluate the performance of participating systems are the following:

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- Recall (Sensitivity) is the ratio of correctly predicted positive observations to the all observations in actual class - positive.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F_1 = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \tag{3}$$

4.2 Subtask 2: Ranking

Systems' predictions will be evaluated against the gold truth using Spearman Correlation. Spearman Correlation measures the rank correlation of two variables X and Y:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the ranks of each observation and n is the number of observations.

The script for the evaluation will be provided at

`https://github.com/wic-ita/data`

4.3 Baselines

We provide the same baseline proposed by Raganato et al. (2020). The baseline exploits models based on the BERT architecture (Devlin et al., 2019) for encoding the target sub-words. The encoded representations are concatenated and fed into a logistic classifier. In cases where the target word is split into multiple sub-tokens, the first sub-token is considered. We set the learning rate to 1e-5 and weight decay to 0. The best checkpoint over the ten epochs is selected using the development data. Differently from Raganato et al. (2020). We use as pre-trained model XLM-RoBERTa (Conneau et al., 2020). *Baseline 1* (Subtask 1) is trained to assign the correct class against the gold truth. *Baseline 2* (Subtask 2) is trained to minimise the difference between the model prediction and the gold score computing the mean squared error.

		Class 0	Class 1
Training		806	1,999
Development	IV	167	236
	OOV	83	14

Table 2: Subtask 1: Dataset statistics. IV: In-Vocabulary, OOV: Out-Of-Vocabulary

5 Provided data

We provide all datasets in the JSON Lines text format containing one example for each line as sketched in Listing 1 for Subtask1 and Listing 2 for Subtask2. Datasets are available at

`https://github.com/wic-ita/data`

5.1 Subtask 1: Binary Classification

We provide two datasets for model development:

- The *train.jsonl* which consists of 2,805 training examples. This dataset should be employed to train the model
- The *dev.jsonl* which consists of 500 training examples. This dataset should be employed to evaluate the model in the training phase, e.g., tune hyper-parameters

The training dataset (*train.jsonl*) is highly unbalanced, consisting of about 71.27% of positive and 28.73% of negative examples. At the same time, we provide a balanced development set (*development.jsonl*) consisting of 50% positive and 50% of negative examples. Further, the dev set includes 250 examples where the target word is out of the vocabulary, i.e., the target word never appears in the training set. For each In-Vocabulary target word of the development set, at least one positive and one negative example are provided in the training set. Overall statistics are reported in Table 2.

5.2 Subtask 2: Ranking

We provide four datasets for model development:

- The *train_agr.jsonl* which consists of 2,805 training examples for which the two annotators agree[1]
- The *train_dis.jsonl* which consists of 1,015 training examples for which the two annotators disagree[2]
- The *train.jsonl* which consists of 3,820 training examples. This dataset is the union of the *train_agr.jsonl* and *train_dis.jsonl* datasets
- The *dev.jsonl* which consists of 500 training examples

Both *train_agr.jsonl* and *dev.jsonl* contain the same examples of the training and development set of Subtask 1.

6 How to submit your runs

6.1 Submission

Each participant can submit at least **three** runs. For each run, a short description of the system and the list of data and resources used must be submitted by filling out the form at (You have to log in with a Gmail account):[3]

`https://forms.gle/2U5sJnpwFJF7F8Pg6`

The form requires to provide the following information:

[1] See Section 3.1

[2] See Section 3.1

[3] *Currently, the form is closed, the submission will open during the Evaluation Window (7th – 14th May 2023).*

- Name of the team
- Name of the run
- Zipped file containing at least two files (*description.txt*, *binary.jsonl* **or/and** *ranking.jsonl* **or/and** *binary_eng.jsonl* **or/and** *ranking_eng.jsonl*). The run can refer to only one of the two subtasks.

The *binary.jsonl* is a JSON Lines text file containing for each row a Json line with the example ID and the predicted label (as in Listing 3).

```
{"id": "ricevere.verb.1", "label": 1}
{"id": "cannone.noun.10", "label": 0}
```

Listing 3: Result format for Subtask 1.

The *binary_eng.jsonl* is a JSON Lines text file containing for each row a JSON with the example ID and the predicted label (as in Listing 4).

```
{"id": "receive_ricevere.verb.1", "label": 1}
{"id": "bomb_bomba.noun.10", "label": 0}
```

Listing 4: Result format for Subtask 1 Cross-lingual.

The *ranking.jsonl* is a JSON Lines text file containing for each row a JSON with the example ID and the predicted score (as in Listing 5).

```
{"id": "ricevere.verb.1", "label": 3.342}
{"id": "cannone.noun.10", "label": 1.002}
```

Listing 5: Result format for Subtask 2.

The *ranking_eng.jsonl* is a JSON Lines text file containing for each row a JSON with the example ID and the predicted score (as in Listing 6).

```
{"id": "receive_ricevere.verb.1", "label": 3.342}
{"id": "bomb_bomba.noun.10", "label": 1.002}
```

Listing 6: Result format for Subtask 2 Cross-lingual.

The *description.txt* file contains information about the submission, such as the model used, the parameters, and other relevant stuff. Please, be sure to properly include in the description all the information necessary to track the submitted system.

An example of submission is available at:

`https://github.com/wic-ita/data/blob/main/xlm_finetuned.zip`

Please, contact us if you experience any issues with the submission at `wicita.evalita@gmail.com`.

References

Carlos Santos Armendariz, Matthew Purver, Matej Ulcar, Senja Pollak, Nikola Ljubesic, and Mark Granroth-Wilding. 2020. CoSimLex: A Resource for Evaluating Graded Word Similarity in Context. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5878–5886. European Language Resources Association.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4330–4338. ijcai.org.

Susan Windisch Brown. 2008. Choosing sense distinctions for WSD: psycholinguistic evidence. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers*, pages 249–252. The Association for Computer Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulic, and Anna Korhonen. 2021. AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7151–7162. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 24–36. Association for Computational Linguistics.

George A. Miller. 1992. WORDNET: a lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 216–225. The Association for Computer Linguistics.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1267–1273. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, José Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7193–7206. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 169–174. Association for Computational Linguistics.