WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task

Pierluigi Cassotti[1,†], Lucia Siciliani[1,*,†], Lucia C. Passaro[2,†], Maristella Gatto[3,†] and Pierpaolo Basile[1,†]

[1]Department of Computer Science, University of Bari, Italy
[2]Department of Computer Science, University of Pisa, Italy
[3]Dipartimento di Ricerca e Innovazione Umanistica, University of Bari, Italy

Abstract

WiC-ita is a shared task proposed at the EVALITA 2023 campaign. The task focuses on the meaning of words in specific contexts and has been modelled as both a binary classification and a ranking problem. Overall, 4 groups took part in both subtasks, with 9 different runs. In this report, we describe how the task was set up, we report the system results, and we discuss them.

Keywords

Word in Context, Lexical Semantics, Evaluation, Dataset

1. Introduction and motivation

Word Sense Disambiguation [1] is a Natural Language Processing task with a long history and is extremely interesting for the Computational Linguistics community. In Word Sense Disambiguation (WSD), the goal is to disambiguate each word occurrence assigning to it the correct sense from a predefined sense inventory, such as WordNet [2]. 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 [3].

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 predefined 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.

In general, the WiC 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 [4], where models improve their performance by training in other languages. Several initiatives have been proposed throughout the years: the first one [3] 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 [5] dataset which tried to tackle this issue by taking into account a total of 15 languages. Next, the MCL-WiC [4] 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 [6], 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 [7], 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 [8, 9] 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.

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023).
*Corresponding author.
†These authors contributed equally.
pierluigi.cassotti@uniba.it (P. Cassotti); lucia.sicialini@uniba.it (L. Siciliani); lucia.passaro@unipi.it (L. C. Passaro); maristella.gatto@uniba.it (M. Gatto); pierpaolo.basile@uniba.it (P. Basile)

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. With the WiC-ITA task at EVALITA 2023 [10], we aim to fill 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. The task is modelled with two different subtasks, namely a binary classification one (Subtask 1) and a ranking one (Subtask 2). Participants were allowed to participate in one or both of the subtasks. Details and examples of annotation are available on the task website. [1]

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 are:

- 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.

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 $score \in [1, 4]$. A higher score corresponds to a higher degree of semantic similarity.

3. Dataset

The creation of datasets for the WiC task usually relies on using sense inventories, such as WordNet or BabelNet [11]. More specifically, sense inventories are often exploited for selecting target words, which should exhibit polysemia, and for the generation of sentence 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.

	Monolingual	**Cross-lingual**
noun	177	128
verb	65	45
adjective	49	67
adverb	11	20
	302	260

Table 1
Number of target words and number of words pairs per PoS

Differently from previous datasets, for the WiC-ITA task, we relied on sense inventories only for the target words selection stage, while we extract the list of sentence pairs from large unlabelled corpora. Moreover, human annotation is carried out for *all* the sentence pairs, thus making WiC-ITA the largest manually annotated resource for the WiC task.

In addition to this, the WiC-ITA dataset includes both monolingual (Italian) and cross-lingual (English-Italian) data.

The dataset is split into training, development, and test portions. 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.

We create the monolingual datasets by selecting target words based on the number of synsets in WordNet and senses reported in Wiktionary. To achieve this, we generate a list of candidate target words for each part of speech (PoS) using lemmas from both WordNet and Wiktionary. For each lemma w, we calculate the count of WordNet synsets (wns_w) and senses reported Wiktionary (wks_w). We then compute $\min(wns_w, wks_w)$ and order all the target words in descending order.

To construct the cross-lingual dataset, we use the MultiSemCor [12], which is based on SemCor [13], the most extensive and widely used dataset for Word Sense Disambiguation. Specifically, we extracted word pairs (Italian-English) that are frequently translated in SemCor. For these word pairs, we compute the frequency of specific synsets. Then, we took the union of synsets for each word pair and computed the probability distribution over the synsets for both the Italian and English words. The Jensen–Shannon Divergence (*JSD*) is computed for each pair, and the pairs are sorted accordingly in decreasing order.

We sample the top-k words for the monolingual setting and the top-k pair of words for the cross-lingual setting according to the $\min(wns_w, wks_w)$ and the *JSD* respec-

[1] http://wic-ita.github.io/.

First annotator	Second annotator	N. examples	Spearman corr.
Annotator 2	Annotator 1	454	0.63
Annotator 3	Annotator 2	442	0.63
Annotator 5	Annotator 3	445	0.65
Annotator 6	Annotator 4	442	0.55
Annotator 6	Annotator 5	447	0.65
Annotator 7	Annotator 4	440	0.55
Annotator 8	Annotator 1	447	0.67
Annotator 10	Annotator 7	448	0.73
Annotator 10	Annotator 9	444	0.66
Annotator 11	Annotator 8	419	0.61
Annotator 11	Annotator 9	397	0.57
		4825	0.63

Table 2
Monolingual annotation statistics

First annotator	Second annotator	N. examples	Spearman corr.
Annotator 3	Annotator 2	54	0.48
Annotator 4	Annotator 1	46	0.39
Annotator 4	Annotator 3	38	0.38
Annotator 6	Annotator 2	79	0.66
Annotator 7	Annotator 2	52	0.62
Annotator 9	Annotator 5	78	0.60
Annotator 10	Annotator 1	35	0.44
Annotator 10	Annotator 5	51	0.76
Annotator 10	Annotator 8	79	0.54
Annotator 11	Annotator 7	136	0.67
Annotator 11	Annotator 9	45	0.75
		693	0.57

Table 3
Cross-lingual annotation statistics

		Class 0	Class 1
Training		806	1,999
Development	IV	236	167
	OOV	14	83
Test	IV	124	127
	OOV	126	123

Table 4
Subtask 1: number of examples for each class. IV: In-Vocabulary, OOV: Out-Of-Vocabulary

tively. The number of sampled words per PoS tag are reported in Table 1.

The monolingual and the cross-lingual sentence pairs are extracted from the itWaC and ukWaC corpora, both part of the WaCKy project [14, 15]. 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 [16]. itWaC, differently from ukWaC, is lemmatized using Morph-it! and is obtained by crawling web pages under the .it domain.

Each sentence pair extracted from the aforementioned resources has been attributed with the average score assigned by two annotators according to the 4-point relatedness scale, i.e. from 4 (Identical meaning) to 1 (Unrelated), the offsets of the target word on the respective sentences, and the lemma of the target word. Note that we only considered the Italian lemma for the cross-lingual examples, albeit providing the offsets for both languages. The annotation process is carried out using Doccano [17].

Each data point (i.e., sentence pair) is annotated by two independent annotators. Tables 2 and 3[2] show the statistics in terms of number of annotated examples and agreement (computed as the Spearman correlation) for each pair of annotators. In the monolingual setting, the Spearman correlation for the annotations consistently exceeds 0.6, with the exception of two cases. On the other hand, in the cross-lingual setting, the average correlation is lower compared to the correlation obtained in the monolingual setting. However, the correlation between annotators in the cross-lingual scenario is also computed on smaller samples, which can impact the reliability of the computed correlation.

The data points for which at least one of the annotators voted 0 (Cannot decide) were discarded from the official dataset for the sake of simplicity. The score for Subtask 2 is obtained by averaging the scores assigned by the two annotators. The ground truth labels for Subtask 1 *(binary)* were drived from the labels of Subtask 2. Specifically, we considered the data points for which the two annotators agreed, namely the case in which both annotators provided a score in the set $\{1, 2\}$ and the case in which both the annotators provided a score in the set $\{3, 4\}$. In the former case, the example was labelled with 0, while in the latter, it was labelled with 1.

The dataset is available for download on the website of the task.[3] The dataset has been constructed using available corpora. We refer to [14, 15] for the details about copyright and usage. Below, we further describe the details of the two sub-tasks.

3.1. Subtask 1: Ranking

We provide two datasets for model development:

- The training dataset which consists of 2,805 training examples. This dataset should be employed to train the model
- The development dataset 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 monolingual test dataset which consists of 500 examples
- The cross-lingual test dataset which consists of 500 examples

The training dataset is highly unbalanced, consisting of 71.27% of positive and 28.73% of negative examples. At the same time, we provide balanced development and test datasets consisting of 50% positive and 50% of negative examples. For each In-Vocabulary target word of the

[2]The annotator groups for the two tasks are independent.
[3]https://wic-ita.github.io/data/.

development and test datasets, at least one positive and one negative example are provided in the training set. Overall statistics are reported in Table 4.

3.2. Subtask 2: Ranking

We provide four datasets for model development:

- The training dataset which consists of 2,805 training examples for which the two annotators agree (This dataset contains the same examples provided for training in Subtask 1)
- A training dataset which consists of 1,015 training examples for which the two annotators disagree
- An overall training dataset which consists of 3,820 training examples. This dataset include both instances where annotators have reached a consensus and those in disagreement
- The development dataset which consists of 500 examples (This dataset contains the same examples provided for development in Subtask 1)
- The monolingual test dataset which consists of 500 examples (This dataset contains the same examples provided for test in Subtask 1)
- The cross-lingual test dataset which consists of 500 examples (This dataset contains the same examples provided for test in Subtask 1)

4. Evaluation

The ranking of the participating systems is provided according to each subtask and test set. In other words, for each subtask, we provide the evaluation in both the monolingual and the cross-lingual setting.

4.1. Subtask 1 (Binary Classification)

Systems' predictions are evaluated against the ground truth using the macro F1-Score, i.e. we compute the F1-score for each class and we take the average of these scores to obtain the macro F1-score.

4.2. Subtask 2 (Ranking)

Systems' predictions are evaluated against the ground truth using Spearman's rank correlation. It measures the rank correlation of two variables X and Y:

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

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.

4.3. Baseline models

The baseline model for the task has been constructed according to [5]. It exploits the BERT architecture [18] for encoding the target sub-words. To deal with cases in which the target word is split into multiple sub-tokens, the first sub-token is considered. Differently from [5], we use as pre-trained model XLM-RoBERTa [19] and train the baseline to minimise the difference between the model prediction and the gold score computing the mean squared error.

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.

The binary baseline for Subtask 1 applies the threshold $\delta = 2$ to the model predictions to obtain discrete labels.

To ensure fair reproducibility and comparisons, the evaluation scripts are available for download. [4]

5. Participants

Overall, different teams participated in the task with 9 distinct runs. We highlight below the main strategies adopted by the teams to deal with the WiC-ITA tasks.

The **BERT 4EVER** team [5] proposed three variants of a system based on BERT. The strategy behind the first model involves using the Labse pre-trained model to perform matching judgment tasks. It applies four different strategies for encoding and matching the spliced sentences, including the addition of [CLS] vectors and siamese vectors. The output probabilities of the four models are fused, with task 2 treated as a six-classification task. The second model for task 1 uses the bert-base-italian-cased pre-trained model and follows the same encoding and matching strategies as the first model. Again, the output probabilities of the four models are fused. For task 2, the Labse pre-trained model is used, and the strategies are identical to those in Model 1, but the predicted classification results are averaged. Finally, a third variant combines both the bert-base-italian-cased and Labse pre-trained models. It applies the same encoding and matching strategies as the previous models, but this time, the output probabilities of all eight models (four from each pre-trained model) are fused.

The **ExtremITA** team proposed two models fine-tuned on the EVALITA 2023 training data. The first system is based on the Large Language Model from Meta AI (LLaMA), i.e., the Italian version called Camoscio [20]. The model is pre-trained to generate text based on user instructions and fine-tuned on task-specific triples of <task, input, output> derived from the training data of EVALITA

[4] https://github.com/wic-ita/data/blob/main/evaluation.py.

[5] The team did not submit the final report.

2023 challenges. The LoRA technique for training was applied, and the model is further fine-tuned on the EVALITA 2023 training data. The second system is based on the Italian version of T5 (IT5) [21]. It underwent fine-tuning on task-specific input-output pairs derived from the training data of EVALITA 2023 challenges. The phrasal forms from the training data were used to train the model. The details of the models developed by the team are reported in [22].

The **LG** team proposed a single system based on the automatic translation of target words in different languages. Opus-MT models have been used for the translation of data into 21 languages. The words are lemmatized and aligned, and the feature vectors are created from the equivalence of the target lemma in translation. Then SVMs are used for solving tasks. PoS-Tagging and lemmatization of Italian sentences have been performed through TreeTagger[6]. Lemmatization in 21 languages has been roughly performed through Simplemma[7]. The details of the models developed by the team are reported in [23].

The models developed by **The Time-Embedding Travelers** team (afterwards mentioned as TTET) are all based on the XLM-RoBERTa-base architecture. Each model is a straightforward threshold-based classifier that utilises the conditional number of the cosine similarity or distance matrix to make predictions. The embeddings of the target word are extracted from both sentences, and pairwise similarities or distances are calculated. The threshold for classification is tuned by selecting the value that maximizes accuracy on a combined train and dev set. The final threshold for prediction is determined as the average of the threshold values obtained from multiple iterations. Model 1 and Model 2 use the last 4 layers of embeddings, while Model 3 uses embeddings from all 12 layers. The details of the models developed by the team are reported in [24].

6. Results

Table 5 reports the results referred to each subtask. Concerning the first subtask, namely the one in which participants were asked to provide a binary classification, the best results were obtained by the LG and the TTET Teams. Specifically, the LG team was ranked first on the Italian test set, while the TTET was ranked first for the Italian-English test set. The results are reported in Table 5.

With respect to the second subtask, where participants were asked to provide a ranking, the best results were obtained by the baseline for the Italian test set and by the TTET Team for the Italian-English test set. However, none of the proposed systems provided satisfactory results for the Italian test set, but the TTET team was ranked first for the Italian-English test set. The results are reported in Table 5.

[6]https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[7]https://pypi.org/project/simplemma/

Team	Run	it-it	it-en
LG	LG	**0.734**	-
TTET	conditional	0.668	**0.731**
TTET	cond2	0.658	0.715
TTET	cond3	0.625	0.738
extremITA	it5	0.611	0.616
baseline		*0.594*	*0.555*
BERT 4EVER	run 3	0.556	0.492
BERT 4EVER	run 2	0.561	0.521
BERT 4EVER	run 1	0.535	0.493
extremITA	camoscio lora	0.513	0.544

Table 5
Subtask 1 (Binary Classification) Results

Team	Run	it-it	it-en
baseline		***0.569***	*0.406*
TTET	conditional	0.553	0.538
TTET	cond2	0.521	**0.548**
TTET	cond3	0.493	0.533
LG	LG	0.492	-
BERT 4EVER	run 1	0.337	0.159
BERT 4EVER	run 2	0.303	0.15
BERT 4EVER	run 3	-	-
extremITA	camoscio lora	-	-
extremITA	it5	-	-

Table 6
Subtask 2 (Ranking) Results

Table 7 presents detailed results for each system, including the classification of in-vocabulary (IV) and out-of-vocabulary (OOV) words. The aim is to evaluate the system's capability to classify words that were not part of the training data. In this regard, the LG system exhibits the highest performance in Subtask 1 for both IV and OOV words. However, in Subtask 2, only the TTET system surpasses the baseline for OOV words.

Interestingly, the performance on OOV targets shows an overall improvement. We propose that the models may have become overly specialized to the specific distribution of IV word classes during training, resulting in overfitting.

7. Conclusions

The WiC-ITA task was approached by four different teams. The results from the evaluation of four different teams' systems revealed interesting trends. While three of the systems were based on the Transformer architecture, one team developed an SVM classifier based on the output of a Machine Translation system (using the Transformer model). In the binary classification task, the

		Binary			Ranking		
Team	**Run**	**All**	**IV**	**OOV**	**All**	**IV**	**OOV**
BERT 4EVER	run 1	0.535	0.508	0.560	0.337	0.270	0.422
BERT 4EVER	run 2	0.561	0.565	0.557	0.303	0.261	0.365
BERT 4EVER	run 3	0.556	0.531	0.582	-	-	-
LG	LG	**0.734**	**0.693**	**0.775**	0.492	0.425	0.555
TTET	cond2	0.658	0.644	0.672	0.521	0.497	0.523
TTET	cond3	0.625	0.597	0.651	0.493	0.461	0.500
TTET	conditional	0.668	0.651	0.685	0.553	0.485	**0.582**
extremITA	camoscio lora	0.513	0.446	0.575	-	-	-
extremITA	it5	0.611	0.586	0.635	-	-	-
baseline		0.594	0.620	0.566	**0.569**	**0.536**	0.567

Table 7
Detailed results for subtasks and participants.

best-performing systems demonstrated a significant improvement over the baseline by 14 percentage points on the Italian test set and 17 percentage points on the English test set. However, in the ranking task, the baseline system outperformed all the proposed systems for the Italian test set, whereas the proposed systems achieved a notable enhancement of 14 percentage points over the baseline for the Italian-English test set.

For the Italian test set, the best result was achieved by the system based on SVM and Machine Translation. This team submitted results only for the monolingual task. In the English test set, the best result is obtained by the system based on the XLM-RoBERTa-base architecture. It is interesting to underline that the worst performances were obtained by the system that adopts instructions-based fine-tuning of a specific LLM for Italian. On the one hand, these results highlight the effectiveness and potential of the different systems in addressing the classification and ranking tasks for the meaning of words in context. On the other hand, the results of the competition highlight that there is still room for improvement and that the task is still far from the results obtained by similar campaigns in English.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent Trends in Word Sense Disambiguation: A Survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4330–4338. URL: https://doi.org/10.24963/ijcai.2021/593. doi:10.24963/ijcai.2021/593.

[2] G. A. Miller, 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, 1992. URL: https://aclanthology.org/H92-1116/.

[3] M. T. Pilehvar, J. Camacho-Collados, WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, 2019, pp. 1267–1273. URL: https://doi.org/10.18653/v1/n19-1128. doi:10.18653/v1/n19-1128.

[4] F. Martelli, N. Kalach, G. Tola, R. Navigli, SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC), in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021, Association for Computational Linguistics, 2021, pp. 24–36. URL: https://doi.org/10.18653/v1/2021.semeval-1.3. doi:10.18653/v1/2021.semeval-1.3.

[5] A. Raganato, T. Pasini, J. Camacho-Collados, M. T. Pilehvar, XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 7193–7206. URL:

https://doi.org/10.18653/v1/2020.emnlp-main.584. doi:10.18653/v1/2020.emnlp-main.584.

[6] Q. Liu, E. M. Ponti, D. McCarthy, I. Vulic, A. Korhonen, AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 7151–7162. URL: https://doi.org/10.18653/v1/2021.emnlp-main.571. doi:10.18653/v1/2021.emnlp-main.571.

[7] C. S. Armendariz, M. Purver, M. Ulcar, S. Pollak, N. Ljubesic, M. Granroth-Wilding, CoSimLex: A Resource for Evaluating Graded Word Similarity in Context, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 5878–5886. URL: https://aclanthology.org/2020.lrec-1.720/.

[8] D. Schlechtweg, S. S. im Walde, S. Eckmann, Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change, in: M. A. Walker, H. Ji, A. Stent (Eds.), 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), Association for Computational Linguistics, 2018, pp. 169–174. URL: https://doi.org/10.18653/v1/n18-2027. doi:10.18653/v1/n18-2027.

[9] S. W. Brown, 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, The Association for Computer Linguistics, 2008, pp. 249–252. URL: https://aclanthology.org/P08-2063/.

[10] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] R. Navigli, S. P. Ponzetto, BabelNet: Building a Very Large Multilingual Semantic Network, in: J. Hajic, S. Carberry, S. Clark (Eds.), ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, The Association for Computer Linguistics, 2010, pp. 216–225. URL: https://aclanthology.org/P10-1023/.

[12] L. Bentivogli, E. Pianta, Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus, Natural Language Engineering 11 (2005) 247–261. doi:10.1017/S1351324905003839.

[13] G. A. Miller, C. Leacock, R. Tengi, R. Bunker, A Semantic Concordance, in: Human Language Technology: Proc. of a Workshop Held at Plainsboro, New Jersey, USA, March 21-24, 1993, Morgan Kaufmann, 1993. URL: https://aclanthology.org/H93-1061/.

[14] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[15] M. Baroni, A. Kilgarriff, Large linguistically-processed web corpora for multiple languages, in: EACL'06: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations; 2006 Apr 5-6; Trento, Italy. Stroudsburg (PA): Association for Computational Linguistics; 2006. p. 87-90, ACL (Association for Computational Linguistics), 2006.

[16] H. Schmid, Probabilistic part-of speech tagging using decision trees, in: New methods in language processing, 2013, p. 154.

[17] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.),

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[20] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[21] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[22] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita@evalita2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[23] L. Gregori, Lg at wic-ita: Exploring the relation between semantic shifts and equivalences in translation, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] F. Periti, H. Dubossarsky, The time-embedding travelers@wic-ita, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.