# SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC)

**Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome, Italy
first.lastname@uniroma1.it

## Abstract

In this paper, we introduce the first SemEval task on Multilingual and Cross-Lingual Word-in-Context (MCL-WiC) disambiguation. This task allows the largely under-investigated inherent ability of systems to discriminate between word senses within and across languages to be evaluated, dropping the requirement of a fixed sense inventory. Framed as a binary classification, our task is divided into two parts. In the multilingual sub-task, participating systems are required to determine whether two target words, each occurring in a different context within the same language, express the same meaning or not. Instead, in the cross-lingual part, systems are asked to perform the task in a cross-lingual scenario, in which the two target words and their corresponding contexts are provided in two different languages. We illustrate our task, as well as the construction of our manually-created dataset including five languages, namely Arabic, Chinese, English, French and Russian, and the results of the participating systems. Datasets and results are available at: https://github.com/SapienzaNLP/mcl-wic.

## 1 Introduction

During recent decades, the field of Natural Language Processing (NLP) has witnessed the development of an increasing number of neural approaches to representing words and their meanings. Word embeddings encode a target word type with one single vector based on co-occurrence information. However, word embeddings conflate different meanings of a single target word into the same representation, thus they fail to capture the polysemous nature of words. To address this limitation, more sophisticated representations such as multi-prototype and contextualized embeddings have been put forward. Multi-prototype embeddings concentrate on the semantics which underlie a target word by clustering occurrences based on their context similarities (Neelakantan et al., 2015; Pelevina et al., 2016). In an effort to exploit the knowledge derived from lexical-knowledge bases, Iacobacci et al. (2015) introduced a new approach which allows sense representations to be linked to a predefined sense inventory. More recently, contextualized embeddings were proposed. These representations are obtained by means of neural language modeling, e.g. using LSTMs (Melamud et al., 2016) or the Transformer architecture (Devlin et al., 2019; Conneau et al., 2020), and are capable of representing words based on the context in which they occur. Contextualized representations have also been used to obtain effective sense embeddings (Loureiro and Jorge, 2019; Scarlini et al., 2020a,b; Calabrese et al., 2020).

Although virtually all the above approaches can be evaluated in downstream applications, the inherent ability of the various embeddings to capture meaning distinctions still remains largely under-investigated. While Word Sense Disambiguation (WSD), i.e. the task of determining the meaning of a word in a given context (Navigli, 2009), has long explored the aforementioned ability, the task does not make it easy to test approaches that are not explicitly linked to existing sense inventories, such as WordNet (Miller et al., 1990) and BabelNet (Navigli and Ponzetto, 2010). This has two major drawbacks. First, sense inventories are not always available, especially for rare languages. Second, such requirement limits the evaluation of word and sense representations which are not bound to a sense inventory. To tackle this limitation, some benchmarks have recently been proposed. The CoSimLex dataset (Armendariz et al.) and the related SemEval-2020 Task 3 (Armendariz et al., 2020) focus on evaluating the similarity of word pairs which occur in the same context. More recently, the Word-in-Context (WiC) task (Pilehvar

and Camacho-Collados, 2019), included in the SuperGLUE benchmark for Natural Language Understanding (NLU) systems (Wang et al., 2019) and its multilingual extension XL-WiC (Raganato et al., 2020), require systems to determine whether a word occurring in two different sentences is used with the same meaning, without relying on a predefined sense inventory. For instance, given the following sentence pair:

- the *mouse* eats the cheese,
- click the right *mouse* button,

the ideal system should establish that the target word *mouse* is used with two different meanings.

Despite the steps forward made in this promising research direction, existing benchmarks suffer from the following shortcomings: i) they are mostly automatically retrieved; ii) they do not enable cross-lingual evaluation scenarios in which systems are tested in different languages at the same time; iii) they do not cover all parts of speech.

In order to address the aforementioned drawbacks, we propose the first SemEval task on Multilingual and Cross-Lingual Word-in-Context (MCL-WiC) disambiguation and present the first entirely manually-annotated dataset for the task. Importantly, MCL-WiC enables new cross-lingual evaluation scenarios covering all parts of speech, as well as a wide range of domains and genres. The dataset is available in five European and non-European languages, i.e. Arabic (Ar), Chinese (Zh), English (En), French (Fr) and Russian (Ru).

## 2 Related Work

Several different tasks have been put forward which go beyond traditional WSD and drop the requirement of fixed sense inventories. Among the first alternatives we cite monolingual and cross-lingual Lexical Substitution (McCarthy and Navigli, 2007; Mihalcea et al., 2010). Word-in-context similarity has also been proposed as a way to capture the dynamic nature of word meanings: the Stanford Contextual Word Similarities (SCWS) dataset, proposed by Huang et al. (2012), contains human judgements on pairs of words in context. Along these same lines, Armendariz et al. introduced CoSimLex, a dataset designed to evaluate the ability of models to capture word similarity judgements provided by humans.

More recently, Pilehvar and Camacho-Collados (2019) presented the Word-in-Context (WiC) dataset. Framed as a binary classification task, WiC is a benchmark for the evaluation of context-dependent embeddings. However, WiC covers only one language, i.e. English, and two parts of speech, namely nouns and verbs. To enable evaluation in languages other than English, Raganato et al. (2020) proposed XL-WiC, an extension of the WiC dataset which covers different European and non-European languages, thus allowing for zero-shot settings. Despite their effectiveness, both the WiC and XL-WiC datasets are not manually created and do not cover all parts of speech. Moreover, they do not consider cross-lingual evaluation scenarios in which systems are tested in more than one language at the same time, thus highlighting the need for a new evaluation benchmark.

| MCL-WiC | | | | |
|---|---|---|---|---|
| Sub-task | Dataset | Train | Dev | Test |
| Multilingual | Ar-Ar | - | 500 | 500 |
| | En-En | 4000 | 500 | 500 |
| | Fr-Fr | - | 500 | 500 |
| | Ru-Ru | - | 500 | 500 |
| | Zh-Zh | - | 500 | 500 |
| Cross-lingual | En-Ar | - | - | 500 |
| | En-Fr | - | - | 500 |
| | En-Ru | - | - | 500 |
| | En-Zh | - | - | 500 |

Table 1: The MCL-WiC dataset: number of unique lexemes divided by sub-task and dataset. The second column (Dataset) indicates the available language combination.

## 3 The Multilingual and Cross-lingual Word-in-Context Task

In this Section, we present our SemEval task and describe a new dataset called Multilingual and Cross-lingual Word-in-Context (MCL-WiC). The task is divided into a multilingual and a cross-lingual sub-task, each containing different datasets divided according to language combination. Each dataset instance is focused on a given lexeme[1] and is composed of a unique ID, a target lemma, its part of speech, two sentential contexts in which the target lemma occurs, and positional indices for retrieving the target words in each sentence. In both sub-tasks, for each lexeme, we provide two

[1]Each lexeme corresponds to a lemma and its part of speech.

| ID | Lemma | POS | Start | End | Sentence |
|---|---|---|---|---|---|
| training.en-en.624 | leave | VERB | 47 | 51 | As mentioned, it was clear that people usually **left** their homelands in search of a better life. |
| | | | 13 | 17 | It should be **left** entirely to the parties to a dispute to choose the modalities of settlement they deemed most appropriate. |
| training.en-en.625 | leave | VERB | 47 | 51 | As mentioned, it was clear that people usually **left** heir homelands in search of a better life. |
| | | | 80 | 87 | However, no hasty conclusion should be drawn that the Republic of Macedonia was **leaving** no room for future improvement. |

Table 2: Excerpt from the multilingual dataset (En-En): two sentence pairs sharing the same first sentence are shown, with the target word occurrence in bold type.

| ID | Tag |
|---|---|
| training.en-en.624 | F |
| training.en-en.625 | F |

Table 3: Example of gold file.

different instances which share one sentence[2]. We provide training and development data only for the multilingual sub-task, whereas test data is provided for both sub-tasks. While training data is produced only in English, both the development and the test data are available in other languages as well. Table 1 provides an overview of the composition of the dataset, which we detail further in the remainder of this paper. Compared to existing datasets, MCL-WiC makes it possible to perform a thorough, high-quality evaluation of a multitude of approaches, ranging from architectures based on pre-trained language models to traditional WSD systems.

In the following, we introduce the multilingual and cross-lingual sub-tasks. Then, we describe the data sources, the selection of the target lexemes and sentence pairs and, finally, the annotation process.

### 3.1 Multilingual sub-task

This sub-task allows systems to be evaluated in a scenario in which only one language at a time is considered. To this end, we manually select sentence pairs in the following language combinations: Ar-Ar, En-En, Fr-Fr, Ru-Ru and Zh-Zh. The multilingual sub-task includes training, development and test splits as reported in Table 1 (top). The training data, available only in English, contains 4000 unique lexemes and 8000 sentence pairs. Instead, both the development and test data splits include 500 unique lexemes and 1000 sentence pairs for each of the aforementioned language combinations. To avoid any bias, each dataset contains a balanced number of tags, i.e. 50% True (T) and 50% False (F).

In Table 2,[3] we report two instances derived from En-En, which share the first sentence. Given the target lemma *leave*, its part of speech (verb) and two sentences in which two occurrences of *leave* are contained, participating systems are required to determine whether the target occurrences (shown in bold type in the Table) share the same meaning (T) or not (F). Since the senses of the target occurrences differ in both sentence pairs, they are both tagged with F in the gold file, as shown in Table 3. Note that, in MCL-WiC, target occurrences can be inflected forms of the target lemma.

### 3.2 Cross-lingual sub-task

The cross-lingual sub-task allows systems to be tested and compared in a cross-lingual scenario. Here, sentence pairs are composed of a sentence in English and a sentence in one of the other MCL-WiC languages, including the following language combinations: En-Ar, En-Fr, En-Ru and En-Zh. It is worth mentioning that, in contrast to past efforts, all sentences are manually selected and annotated, and that Arabic and Russian are included in a Word-in-Context dataset for the first time.

We report two cross-lingual instances (sentence pairs) in Table 4 for the En-Ru language combi-

[2]To speed up the annotation process, for each lexeme, we selected a fixed sentence and annotated two other sentences so as to obtain two instances.

[3]Due to space limits we removed some words from the sentences reported in Table 2 and 4.

| ID | Lemma | POS | Start | End | Sentence |
|---|---|---|---|---|---|
| test.en-ru.18 | light | NOUN | 46 | 51 | Using a technique for concentrating the solar **light**, resulted in an overall efficiency of 20%. |
| | | | 39 | 50 | Каждый представитель может выступать в **зависимости** от полученных указаний. |
| test.en-ru.19 | light | NOUN | 46 | 51 | Using a technique for concentrating the solar **light**, resulted in an overall efficiency of 20%. |
| | | | 2 | 8 | С **учетом** работы, оратор считает целесообразным вновь изложить принципы. |

Table 4: Excerpt from the cross-lingual dataset (En-Ru): two sentence pairs sharing the same first sentence are shown, with the target word occurrence in bold type.

nation, which share the first sentence. Given the English lemma *light*, its part of speech (noun), and two sentences, one in English where *light* occurs and one in Russian where a translation of *light* appears, participants are asked to determine whether the target occurrence (in bold in the Table) of *light* and its translations into Russian зависимости and учетом share the same meaning or not. Importantly, translations are allowed to be multi-word expressions and periphrases.

The cross-lingual sub-task comprises test data only and includes 500 unique English lexemes and 1000 sentence pairs for each language combination as reported in Table 1 (bottom). Note that, in this case, all cross-lingual datasets share the same English target lexemes. Similarly to its multilingual counterpart, the data in this sub-task contains a balanced number of T (50%) and F (50%) tags.

### 3.3 Selection of the data and annotation

**Sources of the data** In order to construct MCL-WiC, we leveraged three resources. First, we used the BabelNet[4] multilingual semantic network (Navigli and Ponzetto, 2010) to obtain a set of lexemes in all languages of interest. Subsequently, we extracted sentence pairs containing occurrences of such lexemes from two corpora, namely the United Nations Parallel Corpus (Ziemski et al., 2016, UNPC)[5] and Wikipedia[6]. UNPC is a collection of official records and parliamentary documents of the United Nations available in the six UN languages[7], whereas Wikipedia is a wide-coverage multilingual collaborative encyclopedia. These corpora were selected due to their wide coverage in terms of domains and languages. In fact, such heterogeneity allowed for the creation of a new competitive benchmark capable of evaluating the generalization ability of a system in discriminating senses in different domains and across languages. With this aim in view, we derived 50% of the selected sentence pairs from UNPC and the remaining 50% from Wikipedia.

**Selection of lexemes** Starting from BabelNet, we extracted a set of 5250 unique ambiguous lexemes in English and 1000 unique lexemes for each of the following languages: Arabic, Chinese, French and Russian. The selected pairs in English were distributed as follows: 4000 for the training data, 500 for the development data and 750 for the test data (500 for the multilingual sub-task and 250 for the cross-lingual sub-task[8]; we enriched the latter with additional 250 pairs derived from the multilingual test data). Instead, the selected pairs in languages other than English were included in the multilingual sub-task only and distributed as follows: 500 for the development data and 500 for the test data. We selected the target lexemes starting from basic vocabulary words and such that they had at least three senses in BabelNet. A key goal was to cover all open-class parts of speech, namely nouns, verbs, adjectives and adverbs, whose distribution in MCL-WiC is shown in Table 5. The target lexemes were chosen so as to avoid phrasal verbs and multi-word expressions.

**Selection and annotation of sentence pairs** For each of the target lexemes, we annotated two sentence pairs from either UNPC or Wikipedia. All selected sentences were well-formatted and, most importantly, provided a sufficient semantic context to determine the meaning of the target occurrences

[4] https://babelnet.org/
[5] https://conferences.unite.un.org/uncorpus/
[6] https://wikipedia.org
[7] Arabic, Chinese, English, French, Spanish and Russian.

[8] We recall that, in the cross-lingual sub-task, the target lexemes are provided in English and shared across all datasets.

| | En-En | | | Ar-Ar | | Fr-Fr | | Ru-Ru | | Zh-Zh | | En-* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Dev | Test | Dev | Test | Dev | Test | Dev | Test | Dev | Test | Test |
| NOUN | 4124 | 582 | 528 | 490 | 494 | 548 | 514 | 572 | 582 | 520 | 554 | 458 |
| VERB | 2270 | 246 | 298 | 428 | 398 | 262 | 272 | 352 | 372 | 330 | 364 | 320 |
| ADJ | 1430 | 158 | 144 | 72 | 98 | 156 | 184 | 54 | 30 | 122 | 62 | 178 |
| ADV | 176 | 14 | 30 | 10 | 10 | 34 | 30 | 22 | 16 | 28 | 20 | 44 |

Table 5: Part-of-speech distribution in MCL-WiC. * indicates all languages supported in MCL-WiC other than English.

unequivocally. Subsequently, each sentence pair was associated with a tag, depending on whether the target words in the two contexts are used with the same meaning (T) or not (F). To perform both the selection of the data as well as the annotation, we employed eight annotators with a high level of education and linguistic proficiency in the corresponding language; the annotation work required approximately six months. Importantly, all annotators followed specific criteria which we describe in the following paragraph.

**Annotation criteria** We provided each annotator with general annotation guidelines. Besides general criteria, each annotation team[9] established ad-hoc guidelines for specific linguistic issues, some of which will be briefly illustrated in Section 4, below.

General annotation criteria can be broadly divided into grammatical and lexicographic-semantic criteria. The former refer to the format and the grammatical correctness of the sentences to be selected: annotators were asked to choose well-written sentences only, i.e. sentences with a clear structure, ending with a full stop and containing a main clause. Instead, lexicographic-semantic criteria refer to the attribution of the labels. To determine whether two occurrences were used with the same meaning or not, annotators were asked to use multiple reputable dictionaries (e.g. for English we used the Merriam-Webster, Oxford Dictionary of English and English Collins dictionaries). Moreover, to avoid misperceptions in the same-sense tagging annotations, we asked annotators to justify their choices by providing substitutes for the target occurrences with synonyms, hypernyms, paraphrases or the like. Contrary to what was done in WiC and XL-WiC, we argue that, for the purposes of this task, annotating according to lexicographic motivations, i.e. by using reliable dictionaries, contributes significantly to minimizing the impact of subjectivity, thus producing more adequate and consistent data. Finally, lexicographic-semantic criteria also provided concrete indications and examples regarding the attribution of tags. For instance, T was used if and only if the two target occurrences were used with exactly the same meaning or, in other words, if, using a dictionary, the definition of the two target words was the same.

**Inter-annotator agreement** In order to determine the degree of uncertainty encountered during the annotation process, we computed the inter-annotator agreement. To this end, we randomly selected a sample of 500 sentence pairs from each of the En-En and Ru-Ru multilingual datasets, and 200 sentence pairs from the En-Ar and En-Zh cross-lingual datasets. Validators were provided with the same guidelines used during the annotation process. We calculated the agreement between two different annotators using the Cohen's kappa, obtaining $\kappa$=0.968 in En-En, 0.952 in Ru-Ru, 0.94 in En-Ar and 0.91 in En-Zh, which is interpreted as almost perfect agreement.

**Data format** For each sub-task, we provide two types of file (.data and .gold) in JSON format. The .data files contain the following information: a unique ID, the lemma, its part of speech, the two sentences and the positional indices to identify the target occurrences to be considered (see Tables 2 and 4). Instead, the .gold files include the gold answers, i.e. the corresponding ID and tag, as shown in Table 3.

## 4 Linguistic Issues

In this section, we describe interesting language-specific issues which required additional guidelines. Due to space limits, we focus on languages which do not use the Latin alphabet, i.e. Arabic, Chinese and Russian, illustrating only the most significant issues encountered.

[9] An annotation team is made up of annotators working on the same language.

**Arabic** From a WSD perspective, compared to other languages, written Arabic poses bigger challenges due to the omission of vocalization, which increases the degree of semantic ambiguity. In fact, the vocalization, expressed by diacritics placed above or below consonants, contributes significantly to determining the right interpretation and thus the meaning of words. For instance, the unvocalized word form *b-r-d* could be interpreted as *bard* ("cold"), *burd* ("garment") or *barad* ("hail"). Of course, in Arabic, polysemy also affects vocalized words, which can have multiple meanings, e.g. *ummiyy* means "maternal", but also "illiterate". For the purposes of MCL-WiC, we chose to keep the sentences as they are found in UNPC and Wikipedia, i.e. unvocalized in the vast majority of cases, while – instead – providing the target lemmas in the vocalized form. This was done in order to avoid lexical ambiguity deriving from lemmas which share the same word form but are vocalized in a different way. Furthermore, this choice facilitated the selection and annotation of sentence pairs in which a given target lemma occurs.

**Chinese** Since Chinese does not adopt an alphabet, the semantic ambiguity that can be found in English homographs is basically lost. In Chinese, if two unrelated words are pronounced in the same way, such as "plane" (the airplane) and "plane" (the surface), they are not usually written in the same way. By way of illustration, 沉默, meaning "silent; to be silent" and 沉没, "to sink", are both pronounced as *chénmò*, but, because they are written with different characters, they cannot be considered ambiguous words. Analogously, some characters have an extremely high semantic ambiguity themselves, but since they appear most frequently in polysyllabic words, their ambiguity is lost. For example, the character *guǒ* 果 has at least two meanings, "fruit" and "result", but this character almost never stands as a word on its own in contemporary Chinese. In the current lexicon most of the Chinese words are composed of two or more characters; when it appears in actual texts, *guǒ* is almost always connected to other characters, and the word thus formed is no longer semantically ambiguous. Finally, similarly to the cross-lingual sub-task, some ambiguity had to be discarded in translation, as in the case of Chinese classifiers which have a marked potential for semantic ambiguity. For example, *dào* 道 is, among others, the classifier for long and narrow objects, as in *yī dào hé* 一道河, a river (one+classifier+river), or for doors, walls and similar objects with an entry and an exit, as in *yī dào mén* 一道门, a door (one+classifier+door). However, since classifiers are virtually absent in European languages, they could not be applied in the cross-lingual sub-task and were discarded.

**Russian** A noteworthy issue encountered by Russian annotators concerned the verbal aspects which can be viewed as one of the most challenging features of the Russian language especially for L2-learners[10] with no Slavic background. In Russian, a verb can be perfective, imperfective or both. Normally, a perfective verb has one or more imperfective counterparts and vice versa. Broadly speaking, perfective verbs are typically used to express non-repetitive actions completed in the past, or actions which will certainly be carried out in the future, and also in general for past or future actions for which the speaker intends to emphasize the result that was or will be achieved. Conversely, imperfective verbs are used to express actions which are incomplete, habitual, in progress, or actions for which the speaker does not stress the result to be attained. In MCL-WiC, given a verbal target lexeme, we decided to choose sentences in which the target words occurring in the selected sentences and the target lemma shared the same aspect. In fact, in Russian, although pairs of perfective and imperfective verbs such as делать, сделать (to do) or спрашивать, спосить (to ask) show a high degree of morphological relatedness, they tend to be considered as distinct lemmas.

Another interesting issue regards participles. In some cases, annotators raised issues concerning the part of speech of participles occurring as target words in the selected sentences. In fact, Russian participles derive from verbs, but are declined and can behave as adjectives. Since the target lexemes and the corresponding occurrences must share the same part of speech, we decided to discard sentences in which the part of speech of the target words could not be determined unequivocally.

## 5 Participating Systems

This Section is devoted to the participating systems. First, we briefly describe the rules of the competition. Subsequently, we provide an overview of the data and approaches used by participants. Then, we focus on some of the best-scoring systems and

[10] In language teaching, L2 indicates a language which is not the native language of the speaker.

provide a breakdown of the techniques adopted. We report the three best-performing teams for each sub-task and language combination in Tables 6 and 7. All results are publicly available on the official MCL-WiC page on GitHub[11]. For each winning team, we show only the best performance in the corresponding category.

### 5.1 Rules of the competition

Participants were given no constraints as far as data was concerned; for instance, the development data could be used for training or it was allowed to enrich the provided data by constructing new datasets in an automatic or semi-automatic fashion. Furthermore, we allowed more than one participant for each team. Participating teams could upload up to five submissions, each including up to 9 language combinations for the two sub-tasks.

### 5.2 Data

**Multilingual sub-task** As far as English is concerned, the majority of participating systems used the MCL-WiC training and development data. Some participants also used the data derived from WiC and XL-WiC. Furthermore, automatically-constructed WiC-like datasets were obtained by some participants, starting from semantic resources such as SemCor (Miller et al., 1993), WordNet and the Princeton WordNet Gloss Corpus (PWNG)[12], or by automatically translating available datasets into English. The available data was also enriched via sentence reversal augmentation (given a sentence pair, the two sentences were swapped). In some cases, the development and trial[13] data was used to enrich the training data.

As regards languages other than English, most participants used XL-WiC data, or new training and development datasets were obtained by splitting the MCL-WiC language-specific development data. Alternatively, in zero-shot scenarios, participants trained their models using the English training data. Furthermore, some participants augmented the training and development data by including the trial data. Also in this case, training and development splits were augmented via sentence reversal.

**Cross-lingual sub-task** In the cross-lingual sub-task, most participants used the MCL-WiC English training and development data in zero-shot settings. A smaller group of participants used WiC and XL-WiC data. Some participants created additional training and development data from other resources such as the Open Multilingual WordNet and PWNG. Additional training and development data was produced via Machine Translation.

| Dataset | Team | Score |
|---|---|---|
| **Ar-Ar** | Cam | **84.8** |
| | LIORI | 84.6 |
| | MCL@IITK; DeathwingS | 84.5 |
| **En-En** | MCL@IITK; oyx | **93.3** |
| | zhestyatsky | 92.7 |
| | Cam | 92.5 |
| **Fr-Fr** | MCL@IITK | **87.5** |
| | Cam | 86.5 |
| | LIORI | 86.4 |
| **Ru-Ru** | Cam | **87.4** |
| | LIORI | 86.6 |
| | godzilla | 86.5 |
| **Zh-Zh** | stce | **91.0** |
| | godzilla | 90.8 |
| | PALI | 90.5 |

Table 6: Multilingual section: five best-scoring systems by language combination.

### 5.3 Approaches

**Multilingual sub-task** Most participants used XLM-RoBERTa (Conneau et al., 2020) as pre-trained language model to obtain contextual representations of the target occurrences. Other models frequently used by participants were mBERT, RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2019) and ERNIE (Sun et al., 2020). The majority of participants made use of fine-tuned contextualized embeddings and used logistic regression to perform binary classification. Some participants used ensembles and majority voting.

**Cross-lingual sub-task** Also in this sub-task, XLM-RoBERTa was the most used multilingual language model. Again, the majority of systems obtained contextualized embeddings, passing them to a logistic regression unit. In this case, participants mainly explored zero-shot approaches. Some participants made use of ensembles, adversarial training, pseudo-labelling (Wu and Prasad, 2017) and cross-validation techniques.

[11] https://github.com/SapienzaNLP/mcl-wic

[12] http://wordnetcode.princeton.edu/

[13] As trial data, we provided 4 instances for each sub-task and dataset.

### 5.4 Competition and best-scoring systems

The MCL-WiC competition took place on the CodaLab[14] open Web-based platform and reported 170 participants, out of which 48 uploaded one or more datasets. Overall, 170 submissions were received, the majority of which were focused on the multilingual sub-task and specifically on the En-En dataset. As far as the evaluation metric was concerned, systems were tested using the accuracy score. In what follows, we provide insights regarding the approaches adopted by some of the best-performing participating systems, based on the information we received. For each paragraph we reported the name of the team and of their members in footnote. If the name of the team was not specified, we reported the name of the participant as indicated in Codalab.

**Cam** The Cam team (Yuan and Strohmaier, 2021) made use of the WiC and XL-WiC datasets in addition to the MCL-WIC data. Furthermore, examples from the Sense Complexity Dataset (Strohmaier et al., 2020, SeCoDa) and the Cambridge Advanced Learner's Dictionary (CALD) were extracted. Cam used pre-trained XLM-RoBERTa as underlying language model and added two additional layers on top to perform binary classification with tanh and sigmoid activation, respectively. As input, the following items were concatenated: the representation corresponding to the first token of the sequence, the representations of the target words in both sentences, as well as the absolute difference, cosine similarity and pairwise distance between the two vectors. When the target word was split into multiple sub-tokens, Cam took the average representation rather than the first sub-token. Finally, a two-step training strategy was applied: 1) pre-training the system using out-of-domain data, i.e. WiC, XL-WiC, SeCoDa and CALD; 2) fine-tuning the system on MCL-WiC data.

**godzilla** godzilla enriched the MCL-WiC training data by automatically constructing a dataset starting from WordNet and using Machine Translation. Different types of pre-trained models, such as RoBERTa and XLM-RoBERTa, were adopted. godzilla highlighted the target words by surrounding them with special markings on both sides and appending the target words to the end of each sentence. As architecture, this system used the next sentence prediction models from the hugging face[15] library. Given the strong connection between En-Ar, En-Fr, En-Ru, En-Zh test datasets, pseudo-tagging was used for each language combination. Finally, godzilla applied label smoothing and model merging.

| Dataset | Team | Score |
|---|---|---|
| **En-Ar** | PALI | **89.1** |
| | godzilla | 87.0 |
| | Cam; LIORI | 86.5 |
| **En-Fr** | PALI | **89.1** |
| | godzilla | 87.6 |
| | LIORI | 87.2 |
| **En-Ru** | PALI | **89.4** |
| | godzilla | 88.5 |
| | RyanStark; rxy1212 | 87.3 |
| **En-Zh** | PALI; RyanStark | **91.2** |
| | Cam | 88.8 |
| | MagicPai | 88.6 |

Table 7: Cross-lingual sub-task: three best-scoring systems by language combination.

**LIORI** The LIORI[16] team (Davletov et al., 2021) used the datasets provided in the MCL-WiC competition. Specifically, the training data was enriched with 70% of the development data for Arabic, Chinese, French and Russian, and the whole trial data. Optionally, data augmentation was performed by swapping sentences in each example. LIORI fine-tuned XLM-RoBERTa on a binary classification task and used a 2-layered feed-forward neural network on top of the language model with dropout and the tanh activation function. Sentences in each pair were concatenated by the special token "</s>" and fed to XLM-RoBERTa. As input, the model took the concatenation of the contextualized embeddings of the target words, aggregating over sub-tokens either by max pooling, or just by taking the first sub-token. LIORI used a voting ensemble composed of three models: the first model trained with data augmentation, using the concatenations of the first sub-tokens of the target words; the second trained with data augmentation using max-pooling over sub-tokens; finally, the third trained without data augmentation and using concatenations of the first sub-tokens.

[14] https://competitions.codalab.org/competitions/27054

[15] https://huggingface.co/

[16] The following member of the team LIORI took part in the competition: davletov.

**stce** stce used the MCL-WiC datasets and built additional training data using HowNet (Dong and Dong, 2003). Furthermore, the training data was enriched by pseudo-labelling the test datasets. Data cleaning was performed and target words were surrounded by special markings. The main language model used was XLM-RoBERTa-large. During the training process, dynamic negative sampling was performed for each batch of data fed to the model. At the same time, stce adopted the Fast Gradient Method and added disturbance to the embedding layer to obtain more stable word representations.

**zhestyatsky** Zhestiankin and Ponomareva (2021) augmented the English MCL-WiC training and development data with WiC. Training and development data were split randomly to create a larger training sample which included 97.5% of the data, while leaving only 2.5% for the new development dataset. Then, bert-large-cased embeddings were fine-tuned using AdamW as optimizer with a learning rate equal to 1e-5. Each sentence was split by BertTokenizerFast into 118 tokens maximum. The model was trained for 4.5 epochs and stopped by Early Stopping with patience equal to 2. For each sentence, zhestyatsky took the embeddings of all sub-tokens corresponding to the target word and max pooled them into one embedding. Subsequently, zhestyatsky evaluated the cosine similarity of these embeddings and activated this value through ReLU, with the system predicting True if the output value was above the threshold of 0.5195 obtained using the ROC curve.

**MCL@IITK** First, the MCL@IITK[17] team (Gupta et al., 2021) pre-processed the sentences by adding a signal, either double quotes on both sides of the target word, or the target word itself appended to the end of the sentence. For En-En, MCL@IITK enriched the MCL-WiC training data using sentence reversal augmentation, WiC and SemCor. MCL@IITK obtained embeddings of the target words using the last hidden layer, and passed them to a logistic regression unit. MCL@IITK used ELECTRA, ALBERT, and XLM-RoBERTa as language models and submitted probability sum ensembles. For the non-English multilingual subtask, MCL@IITK used XLM-RoBERTa only and tackled all four language pairs jointly. A 9:1 train-dev split with sentence reversal augmentation was used on the non-English dev data, in addition to En-En train data and XL-WiC with an ensemble model. For the cross-lingual subtask, ELECTRA embeddings were used. The models were trained on partly back-translated En-En train set and validated on back-translated En-En development set.

**PALI** The PALI[18] team (Xie et al., 2021) enriched the MCL-WiC data using WordNet while keeping the original cross-lingual data to maintain the target words in the cross-lingual data. After text pre-processing, task-adaptive pre-training was performed using the MCL-WiC data. The target words were surrounded by special symbols. PALI used XLM-RoBERTa as main language model and took its final output layer, concatenating the [CLS] token with the embeddings of the target occurrences in each sentence pair. To increase the training data, PALI exchanged the order of 20% of the sentence pairs. During training, lookahead (AdamW) was used together with adversarial training implemented by the Fast Gradient Method to obtain more stable word representations. Hyperparameters were tuned through trial-and-errors. The models of stratified 5-fold cross-validation were averaged to yield the final prediction results.

## 6 Baselines

Following Raganato et al. (2020), we used a baseline transformer-based binary classifier. Thus, first, given a sentence pair, a dense representation is obtained for each target occurrence. As indicated in Devlin et al. (2019), in the case that a target occurrence is split into multiple sub-tokens, the first sub-token is selected. The resulting representations are then given as input to a binary classifier implemented following Wang et al. (2019). We selected the Adam optimizer (Kingma and Ba, 2015) with learning rate and weight decay equal to 1e-5 and 0, respectively, and trained for 10 epochs.

We experimented with two different contextualized embedding models: BERT (base-multilingual-cased) and XLM-RoBERTa (base). As for the data, in contrast to most participants, we made use of the data provided for the task only. We used En-En as training and development data for English. As for other language combinations, we trained on En-En and validated both on En-En or and on the other language multilingual development data. Table 8

[17]The following members of the MCL@IITK team took part in the competition: jaymundra, rohangpt and dipakam.

[18]The following members of the PALI team took part in the competition: endworld and xsysigma.

| Model | Ar-Ar | En-En | Fr-Fr | Ru-Ru | Zh-Zh | En-Ar | En-Fr | En-Ru | En-Zh |
|---|---|---|---|---|---|---|---|---|---|
| mBERT$_1$ | **76.2** | 84.0 | **78.7** | 74.5 | 77.5 | 65.9 | 71.6 | 68.2 | **68.9** |
| XLMR-base$_1$ | 75.4 | **86.6** | 77.9 | **76.5** | **78.5** | **67.7** | **71.8** | **74.2** | 66.1 |
| mBERT$_2$ | **76.4** | 84.0 | **78.7** | 74.6 | 76.6 | 62.0 | 69.4 | 66.7 | 64.2 |
| XLMR-base$_2$ | 75.4 | **86.6** | 77.7 | **76.5** | **78.9** | **67.7** | **74.9** | **74.2** | **71.3** |

Table 8: Accuracy of baselines for multilingual and cross-lingual sub-tasks. Columns indicate the test set used. In setting 1, we used the En-En training data and the En-En development data. In setting 2, we used the En-En training data and the corresponding development datasets in languages other than English.

reports the best training results according to the corresponding validation.

## 7 Results and Discussion

In this section, we discuss the results achieved in our competition. Overall, the MCL-WiC dataset allows systems to attain high performances, in the 85-93% accuracy range. This leads us to hypothesize that, in general, systems were able to develop a good ability in capturing sense distinctions without relying on a fixed sense inventory.

When compared to the proposed baselines, we observe that best-performing systems were able to achieve an absolute improvement of up to 27.1 points over the corresponding baselines (e.g. on En-Ar, cf. Tables 7 and 8). Both our baselines and the systems developed by participants confirm that, in this task, XLM-RoBERTa outperforms BERT in most language combinations. The highest score was obtained in En-En, with the best system achieving 93.3% accuracy. Note that our baselines were also able to attain good performances in En-En, i.e. 84.0% using BERT and 86.6% with XLM-RoBERTa, without benefiting from additional training and development data. Interestingly, Chinese was the language which achieved the second-best results, both in Zh-Zh and En-Zh, attaining on average results which were considerably higher. Instead, Arabic seems to have been the most difficult language for participants, especially in Ar-Ar. A reason for this result, deserving further exploration, could lie in morpho-semantic features inherent in Arabic, which we briefly outlined in Section 4.

Zero-shot approaches differ in the performances achieved by participants in the two sub-tasks: in the cross-lingual sub-task participants were able to achieve slightly better performances than those in the multilingual setting, most probably thanks to the presence of English in both the training and the test data, and, more in general, to the availability of English WiC-style datasets which could be used to enrich the already provided data. With the exception of Chinese, instead, on the multilingual sub-task we observe a performance drop between 1.6 and 4.3%.

Finally, we note that performance boosts were observed across the board when using data augmentation, especially by swapping the two sentences within a pair or by coupling the second sentences of two pairs sharing the same first sentence and the same meaning. Another consistent performance increase, observed both in the multilingual and in the cross-lingual sub-task, was obtained when adding a signal on both sides of the target occurrences.

## 8 Conclusions

In this paper, we described the SemEval-2021 Task 2 and introduced Multilingual and Cross-lingual Word-in-Context (MCL-WiC), the first entirely manually-curated WiC-style dataset available in five European and non-European languages, namely Arabic, Chinese, English, French and Russian. MCL-WiC allows the inherent ability of systems to discriminate between word senses within the same language to be tested, and also, interestingly, within cross-lingual scenarios in which a system is evaluated in two languages at the same time, namely English and one of the remaining MCL-WiC languages.

While current Word-in-Context datasets focus primarily on single tokens, as a suggestion for future work we would like to further explore the integration of multi-word expressions and idiomatic phrases into a Word-in-Context task. This would allow us to investigate the intrinsic ability of a system to correctly discriminate the semantics of such linguistic constructs, especially those whose meaning is not compositional, i.e. it cannot be derived by combining the meaning of each of their individual components.

## Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme.

We gratefully thank Luisa Borchio, Ibraam Abdelsayed, Anna Guseva, Zhihao Lyu and Beatrice Buselli for their valuable annotation work.

## References

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 Task 3: Graded Word Similarity in Context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49.

Carlos Santos Armendariz, Matthew Purver, Matej Ulcar, Senja Pollak, Nikola Ljubesic, Marko Robnik-Sikonja, Mark Granroth-Wilding, and Kristiina Vaik. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of The 12th Lan-guage Resources and Evaluation Conference*, page 5878–5886.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. EViLBERT: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 481–487. International Joint Conferences on Artificial Intelligence Organization.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Adis Davletov, Nikolay Arefyev, Denis Gordeev, and Alexey Rey. 2021. LIORI at SemEval-2021 Task 2: Span Prediction and Binary Classification approaches to Word-in-Context Disambiguation. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*, Bangkok, Thailand (online). 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2003. HowNet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Rohan Gupta, Jay Mundra, Deepak Mahajan, and Ashutosh Modi. 2021. MCL@IITK at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation using Augmented Data, Signals, and Transformers. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*, Bangkok, Thailand. Association for Computational Linguistics.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5682–5691. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshopon Semantic Evaluations (SemEval-2007)*, page 48–53.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. Context2Vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61. ACL.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: an online lexical database. *International Journal of Lexicography*, 3(4).

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *CoRR*, abs/1504.06654:1059–1069.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 174–183. Association for Computational Linguistics.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. WiC: the Word-in-Context dataset for evaluating context-sensitive meaning representations. In *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, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proc. of AAAI*, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, page 3528–3539. Association for Computational Linguistics.

David Strohmaier, Sian Gooding, Shiva Taslimipoor, and Ekaterina Kochmar. 2020. SeCoDa: Sense complexity dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5962–5967, Marseille, France. European Language Resources Association.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 3261–3275.

Hao Wu and Saurabh Prasad. 2017. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270.

Shuyi Xie, Jian Ma, Haiqin Yang, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. PALI at SemEval-2021 task 2: Fine-Tune XLM-RoBERTa for Word in Context Disambiguation. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Zheng Yuan and David Strohmaier. 2021. Cambridge at SemEval-2021 Task 2: Neural WiC-Model with Data Augmentation and Exploration of Representation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Boris Zhestiankin and Maria Ponomareva. 2021. Zhestyatsky at SemEval-2021 Task 2: ReLU over Cosine Similarity for BERT Fine-tuning. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.