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 disambiguation (MCL-WiC). 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

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 24–36
Bangkok, Thailand (online), August 5–6, 2021.

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 open-class parts of speech.

In order to address the aforementioned drawbacks, we propose the first SemEval task on Multilingual and Cross-Lingual Word-in-Context disambiguation (MCL-WiC) and present the first entirely manually-annotated dataset for the task. Importantly, MCL-WiC enables new cross-lingual evaluation scenarios covering all open-class 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.

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.

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 open-class 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.

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

[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.

both sub-tasks, for each lexeme, we provide two 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,

[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.