ITA-SENSE - Evaluate LLMs' ability for ITAlian word SENSE disambiguation: A CALAMITA Challenge

Pierpaolo Basile[1,*,†], Elio Musacchio[2,†] and Lucia Siciliani[1,†]

[1] *Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro (ITALY)*

[2] *Italian National PhD Program in Artificial Intelligence, University of Bari Aldo Moro, Bari (ITALY)*

Abstract

The challenge is designed to assess LLMs' abilities in understanding lexical semantics through Word Sense Disambiguation, providing valuable insights into their performance. The idea is to cast the classical Word Sense Disambiguation task in a generative problem following two directions. Our idea is to propose two tasks: (T1) Given a target word and a sentence in which the word occurs, the LLM must generate the correct meaning definition, (T2) Given a target word and a sentence in which the word occurs, the LLM should choose from a predefined set the correct meaning definition. For T1, we compare the generated definition with respect to the correct one taken from a sense inventory, while for T2, a classical accuracy metric is used. In T1, we adopt metrics that measures the quality of the generated definition such as RougeL and the BERTscore. For CALAMITA, we test LLMs using a zero-shot setting.

Keywords

Natural Language Processing, Word Sense Disambiguation, Large Language Models

1. Challenge: Introduction and Motivation

Word Sense Disambiguation (WSD) [1, 2] is a Natural Language Processing task that aims to build a system capable of disambiguating a word occurrence and assigning it the correct sense from an inventory defined a priori, like WordNet [3].

Being a long-standing task in the field of NLP, several techniques have been employed to solve it, reflecting the evolution of advances in machine learning. We can mainly distinguish two main phases. Initially, rule-based systems dominated, followed by knowledge-based methods when digital sense inventories became available. As digital corpora emerged, supervised approaches took advantage of manually annotated data. The vast corpora available on the web and large knowledge graphs further transformed supervised and knowledge-based methods.

The introduction of transformer-based [4] language models marked a new era within the field. These models represent words in context using dense vectors, offering new opportunities for word meaning disambiguation.

Recently, Large Language Models (LLMs) have revolutionized the research in computational linguistics. These models, built on the transformer architecture and trained on massive text datasets, show outstanding capabilities in understanding and generating human-like language. LLMs have demonstrated their ability to solve tasks in zero-shot or few-shot settings, i.e. providing them a prompt without specific training data, though fine-tuning for specific tasks is also possible. Their success suggests an inherent ability to grasp language semantics.

Nevertheless, numerous challenges and issues remain related to LLMs and their actual performance. A key difficulty lies in determining to what extent LLMs are capable of understanding the meaning of a given task rather than merely juxtaposing text coherently. For this reason, tasks like WSD can help to shed light on these issues, as they target specific aspects of natural language. In particular, WSD requires a deep understanding of word meanings in context.

WSD is a task particularly intertwined with the language to be analyzed. In Italian, for example, many words have multiple meanings that can only be adequately understood in context. This is particularly challenging with words with high degree of polysemy. Addressing these ambiguities in Italian makes WSD important for accurately representing the richness of this language. In the past, several evaluation campaigns have been organized such as SensEval and SemEval.

Regarding model performance, we expect LLMs to perform reasonably well at disambiguating common meanings. However, these models may struggle with rare cases (e.g., idiomatic expressions and words belonging to particular domains). We expect fine-tuning on Italian corpora to be essential in developing an LLM capable of addressing this task. The complexity that characterizes Italian morphology and polysemy can be a real challenge

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

[†] These authors contributed equally.

pierpaolo.basile@uniba.it (P. Basile); elio.musacchio@uniba.it (E. Musacchio); lucia.sicialiani@uniba.it (L. Siciliani)

0000-0002-0545-1105 (P. Basile); 0009-0006-9670-9998 (E. Musacchio); 0000-0002-1438-280X (L. Siciliani)

for LLMs unless they are provided extensive language-specific knowledge. For the above reasons, we designed a specific benchmark for CALAMITA [5] to evaluate LLMs' ability in Italian Word Sense Disambiguation.

2. Challenge: Description

Our benchmark aims to measure how an LLM can solve the WSD task for understanding if the model somehow stores knowledge about word meanings. The benchmark is composed of two tasks:

1. Given a sentence and an occurrence of the target word, the model is tested in generating the definition of the word;
2. Given a sentence, the list of possible definitions and an occurrence of the target word, the model is evaluated in selecting the correct definition from the predefined set of possible choices.

Given the same sentence and the target word "squadra", Tables 1 and 2 show the two tasks. Task 1 aims at measuring the LLM ability to generate a definition given a word in a specific context, while Task 2 aims to test the capability of selecting the correct definition from a set of predefined possibilities. The Task 2 is more similar to how the WSD problem is classically formulated in literature, while Task 1 is designed to evaluate the generation capabilities.

Sentence	*"...nonostante l'espulsione di Splitter, la squadra di Ivonic ha mantenuto il ritmo, ha difeso bene..."*
Expected output	*Ritmo di marcia o di corsa.*

Table 1
Example of task 1.

3. Data description

3.1. Origin of data

To create our benchmark, we need an Italian sense-annotated corpus, i.e., a collection of sentences in which each word is tagged with its correct meaning taken from a sense inventory. For this reason, we also require an Italian sense inventory that provides the set of possible meanings for each word.

We use XL-WSD [6] as our sense-annotated corpus. This dataset serves as a cross-lingual evaluation benchmark for the WSD task, featuring sense-annotated development and test sets in 18 languages (including Italian) from six different linguistic families. The sense inventory adopted in XL-WSD is BabelNet [7]. However, not all senses in BabelNet have an Italian gloss. For this reason, we build two versions of the dataset: **without translation** in which we consider only the word occurrences that have Italian glosses in BabelNet, and **with translation** in which English glosses[1] are automatically translated in Italian. For the translation, we use the 1.3B variant of the Meta NLLB-200 model[2].

Sentence	*"...nonostante l'espulsione di Splitter, la squadra di Ivonic ha mantenuto il ritmo, ha difeso bene..."*
Possible choices	1) Rapporto tra due quantità nell'unità di tempo. 2) Ritmo di marcia o di corsa. 3) Il ritmo è una successione di accenti forti e deboli ed eventuali pause, intervallati nel dominio del tempo da pochi decimi di secondo a qualche secondo, che seguono, di solito ma non obbligatoriamente, uno o più modelli ciclici. 4) Alternanza di sillabe di tipi diversi.
Expected output	*2*

Table 2
Example of task 2.

3.2. Data format

We will introduce some formal notations before delving into the description of the benchmark construction. Given a sentence S_k and one of its word occurrences w_i, we define L_i as the list of possible meanings of w_i and $m_j \in L_i$, the meaning assigned to w_i. Each meaning has several glosses, we use $m_j \in L_i$ to refer to it. We need a strategy for building prompts for the two tasks, starting from the Italian sense-annotated corpus and the corresponding sense inventory.

Task 1 aims to assess the LLM's ability to generate an accurate definition of a word in a specific sentence. We create the prompt reported in Table 3 for each sense annotated word occurrence. In the dataset, we also store the correct definition m_j in a field called `output`.

During the construction of the dataset, we need to manage the cases in which a word w_i occurs more than once in the sentence S_k. In these cases, we change the prompt as follows: "*Give a brief definition of the x occurrence of the word "w_i"...*", where $X = \{first, second, third, fourth, fifth\}$ and $x \in X$. We exclude cases where the word occurs more than six times, and we translate the set X according to each language.

[1] The English gloss is always available.
[2] https://huggingface.co/facebook/nllb-200-1.3B

Prompt template (generation)
Give a brief definition of the word "w_i" in the sentence given as input. Generate only the definition. Input: "S_k"
English prompt
Give a brief definition of the word "art" in the sentence given as input. Generate only the definition. Input: "The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world."

Table 3
Prompt for the generation benchmark.

The goal of Task 2 is to evaluate the LLM's ability to select the correct sense from a set of predefined possibilities. In this case, we exploit the list of all possible meanings L_i. In particular, from L_i, we remove all the annotated meanings[3] and obtain the set C_i. Then, we randomly add to C_i one of the correct meanings; in this way, C_i contains only one correct sense. For each occurrence of a sense-annotated word in the corpus, we create the prompt in Table 4. Additionally, we store the identifier (i.e. the option's number) corresponding to the correct answer in a field called `output`.

Prompt template (multiple choice)
Given the word "w_i" in the input sentence, choose the correct meaning from the following: C_i. Generate only the number of the selected option.
English prompt
Given the word "art" in the input sentence, choose the correct meaning from the following: 1) Photographs or other visual representations in a printed publication 2) A superior skill that you can learn by study and practice and observation 3) The products of human creativity; works of art collectively 4) The creation of beautiful or significant things. Generate only the number of the selected option. Input: "The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world."

Table 4
Prompt for the multiple choice benchmark.

We also manage the case where the word w_i occurs more than once by modifying the prompt as in Task 1. Moreover, given that the model is asked to choose among different options in Task 2, we need to manage cases in which the size of C_i is less than two. In these cases, we remove the occurrence from the dataset. Monosemic words are not considered in the construction of both tasks[4].

3.3. Example of prompts used for zero or/and few shots

Our challenge allows only **zero-shot**. Table 5 reports the prompt used in Task 1.

Prompt template (generation)
Fornisci una breve definizione della parola "w_i" nella frase data in input. Genera solo la definizione. Input: "S_k"
Italian prompt
Fornisci una breve definizione della parola "sforzo" nella frase data in input. Genera solo la definizione. Input: "Che sforzo fate per valutare i risultati del vostro programme?"

Table 5
Prompt for the Italian generation task.

Table 6 reports the prompt for Task 2.

Prompt template (multiple choice)
Data la parola "w_i" nella frase in input, scegli il significato corretto tra i seguenti: C_i. Genera solo il numero dell'opzione selezionata. Input: "S_k"
Italian prompt
Data la parola "valutare" nella frase in input, scegli il significato corretto tra i seguenti: 1) Esaminare o ascoltare (prove o un intero caso) per via giudiziaria. 2) Fare la stima commerciale di qlco. 3) Assegnare un valore a. 4) Ritenere dopo valutazione. 5) Apprezzare, tenere in grande stima. 6) Avere una certa opinione di qualcuno. Genera solo il numero dell'opzione selezionata. Input: "Che sforzo fate per valutare i risultati del vostro programme?"

Table 6
Prompt for the Italian multiple choice task.

3.4. Detailed data statistics

Table 3.4 reports the number of instances for each task. We also report different statistics for the dataset without translation and the one with machine translation.

[3] In the sense-annotated corpus, a word occurrence can be annotated with more than one correct meaning.

[4] For Task 1 based on definition generation, it is also possible to consider monosemic words. We exclude this hypothesis since we want to test LLMs in the case of polysemy.

	Task 1	Task 2
without translation	1,673	1,529
with translation	1,888	1,823

Table 7
Dataset statistics.

4. Metrics

The idea is to measure the correspondence between the generated definition and the correct one provided by the sense inventory in Task 1. For Task 2, we want to measure the accuracy in selecting the correct definition from the set of possibilities. For the above reasons, we use three different metrics. For Task 1, we compute F1-RougeL and F1-BERTscore between the reference and generated gloss. For Task 2, we measure the accuracy as the ratio between the correct answers and the number of instances in the dataset.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics that assess the quality of generated texts, particularly summaries, by comparing them with reference texts. Common variants include ROUGE-N, which measures the correspondence of n-grams, ROUGE-L, which considers the longest common sub-sequences, and ROUGE-W, which considers the weight of correspondences. We select the ROUGE-L to measure the lexical correspondence between the generated definition and the correct one. BERTScore relies on pre-trained language models to assess the semantic similarity between the generated and reference definitions, going beyond mere superficial word matching.

If a unique score for Task 1 is necessary, we propose the harmonic mean between RougeL and BERTscore, giving BERTscore double the weight of RougeL. The idea is to give semantic similarity more importance than word matching.

$$\frac{5 * RougeL * BERTscore}{4 * RougeL + BERTscore} \tag{1}$$

We have already performed some evaluations involving several LLMs with a medium number of parameters. Results are reported in Table 4 and show that Llama3.1-8B-Instruct provides the best performance in gloss generation (Task 1), while Gemma2-9B-Instruct achieves the best accuracy.

	Task 1		Task 2
	RougeL	BERTscore	Accuracy
Llama3.1 8B-Instruct	.1363	.6985	.4604
Mistral 7B-Instruct	.0747	.6532	.5324
Gemma2 9B-Instruct	.1221	.6986	.5840

Table 8
Results of several LLMs on our benchmark.

5. Limitations

We cannot guarantee that texts presented in XL-WSD do not occur in the training data of some LLMs. However, even if the model is exposed to textual data from XL-WSD, it does not necessarily mean that it was asked to solve the disambiguation task on such data. A fixed sense inventory may not cover all Italian senses, neologisms, or emerging phrases. However, our benchmark considers only words (and their contexts) annotated according to the sense inventory used in XL-WSD. This ensures that all instances in our benchmark have at least one correct sense in the sense inventory.

6. Ethical issues

No ethical issues are reported in our dataset.

7. Data license and copyright issues

Our data are based on the data license of the XL-WSD from which our benchmark is derived. XL-WSD is distributed under a non-commercial license[5].

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] N. Ide, J. Véronis, Introduction to the special issue on word sense disambiguation: the state of the art, Computational linguistics 24 (1998) 1–40.

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

[3] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[4] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[5] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[5] https://sapienzanlp.github.io/xl-wsd/license/

[6] T. Pasini, A. Raganato, R. Navigli, XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation., in: Proc. of AAAI, 2021.

[7] R. Navigli, S. P. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: J. Hajič, S. Carberry, S. Clark, J. Nivre (Eds.), Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 216–225. URL: https://aclanthology.org/P10-1023.