ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task

Mohammed Khalilia[1] **Sanad Malaysha**[1] **Reem Suwaileh**[2] **Mustafa Jarrar**[1]
Alaa Aljabari[1] **Tamer Elsayed**[3] **Imed Zitouni**[4]
[1]Birzeit University, Palestine [2]Hamad Bin Khalifa University, Qatar
[3]Qatar University, Qatar [4]Google, USA
{mkhalilia, smalaysha, mjarrar, aaljabari} @birzeit.edu
rsuwaileh@hbku.edu.qa telsayed@qu.edu.qa imed.zitouni@gmail.com

Abstract

This paper presents an overview of the Arabic Natural Language Understanding (ArabicNLU 2024) shared task, focusing on two subtasks: Word Sense Disambiguation (WSD) and Location Mention Disambiguation (LMD). The task aimed to evaluate the ability of automated systems to resolve word ambiguity and identify locations mentioned in Arabic text. We provided participants with novel datasets, including a sense-annotated corpus for WSD, called SALMA with approximately 34k annotated tokens, and the IDRISI-DA dataset with 3,893 annotations and 763 unique location mentions. These are challenging tasks. Out of the 38 registered teams, only three teams participated in the final evaluation phase, with the highest accuracy being 77.8% for WSD and the highest MRR@1 being 95.0% for LMD. The shared task not only facilitated the evaluation and comparison of different techniques, but also provided valuable insights and resources for the continued advancement of Arabic NLU technologies.

1 Introduction

Natural Language Understanding (NLU) is a core aspect of Natural Language Processing (NLP), facilitating semantics-based human-machine interactions (Bender and Koller, 2020). One of the key challenges in Arabic is ambiguity, because Arabic exhibits morphological richness, encompassing a complex interplay of roots, stems, and affixes, and rendering words susceptible to multiple interpretations based on their morphology (Jarrar, 2021). Ambiguity in language can lead to misunderstandings, incorrect interpretations, and errors in NLP applications (Maulud et al., 2021). A core NLU task is Word Sense Disambiguation (WSD), and its special case Location Mention Disambiguation (LMD). WSD aims to determine the correct sense of ambiguous words in context (Jarrar et al., 2023c; Al-Hajj and Jarrar, 2021a), while LMD focuses on disambiguating location mentions that are referred to with multiple toponyms, i.e., particular place or location (Suwaileh et al., 2023a).

The Arabic linguistic complexity, coupled with inherent polysemy, underscores the necessity for these lexical disambiguation tasks. They help decipher the intended sense of a word and the targeted entity of a mention within diverse contexts. For instance, in the sentence "شربت من عيون طرابلس وأكلت حلاوة الجبن فيها / I drank from the springs of Tripoli and ate the sweetness of cheese in them" one needs to disambiguate the meaning of two ambiguous words: "عيون / springs," which has 11 related senses, and "طرابلس / Tripoli," which could refer to a location in either Lebanon or Libya.

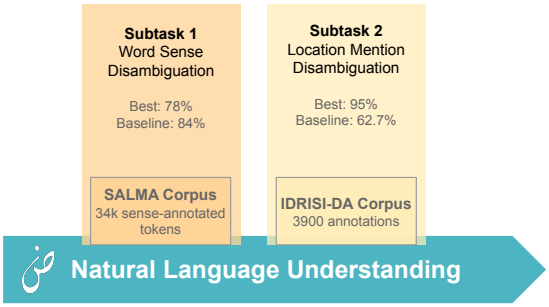

Figure 1: ArabicNLU datasets for WSD and LMD.

WSD is particularly important for tasks like machine translation (Raganato et al., 2020), where it plays a pivotal role in improving accuracy by selecting contextually appropriate translations. Information retrieval systems (Abderrahim and Abderrahim, 2022) also heavily rely on accurate WSD and LMD to ensure search queries yield relevant results, considering the appropriate senses/entities of words within queries. Furthermore, applications such as question answering (Bakari et al., 2021), sentiment analysis (Baiju, 2022), text summarization (Kouris et al., 2021), news analysis (Potey et al., 2020), and semantic search (Modi and Jagtap, 2018) benefit from WSD and LMD. These tasks contribute to a nuanced understanding of Arabic text, enhancing the accuracy and relevance of results across diverse NLP applications. Recently, semantic disambiguation tasks have become integral to addressing hallucinations in Large Language Models (LLMs) (Kritharoula et al., 2023;

Barbon Junior et al., 2024).

While these tasks are extensively researched in well-resourced languages, there is a noticeable lack of focus on Arabic, despite their pivotal role in NLP (Malaysha et al., 2023). This scarcity in Arabic NLP research can largely be attributed to the lack of datasets supporting these essential tasks (Jarrar et al., 2023c). Without sufficient data, researchers face significant obstacles in developing and evaluating models tailored to the complexities of the Arabic language.

To address these challenges and draw attention to the issues faced in Arabic NLU, we organized the first Arabic Natural Language Understanding (ArabicNLU 2024) shared task, focusing on two subtasks: WSD and LMD. We have provided the participating teams with carefully annotated datasets, which are publicly accessible. Specifically, we have provided two manually annotated high-quality datasets for Arabic WSD and LMD. The teams were invited to experiment with diverse deep learning and machine learning methodologies, including, but not limited to, generative approaches, multi-task learning, transfer learning, sequence classification, sequence-to-sequence modeling, and graph models. Despite having 38 registered teams, we received only three submissions, highlighting the challenging and non-trivial nature of the shared task topics.

The remainder of the paper is organized as follows: Section 2 offers a brief literature of Arabic WSD and LMD. Section 3 details the intricacies of the shared task. Section 4 discusses the WSD task, including definition, dataset, baselines, participants' systems, and results. Section 5 discusses the LMD task, including definition, dataset, baselines, participants' systems, and results. Finally, Section 6 concludes the paper.

2 Related Work

NLU enables language models to accurately represent the knowledge embedded in words, which is crucial for core semantic tasks like WSD. WSD remains challenging despite the advancements in deep learning models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT). The challenge extends beyond WSD, involving various disambiguation tasks such as LMD, which aims at disambiguating multiple toponyms for a given location. More complex disambiguation tasks involve intents, anaphora, metaphors, and poetry. Therefore, computational semantics must explore these areas in greater depth, beyond merely considering the Zipfian distribution of words.

2.1 Word Sense Disambiguation (WSD)

Systems Traditionally, rule-based methods (Abeysiriwardana and Sumanathilaka, 2024) dominated utterance ambiguity approaches by leveraging lexical resources such as Qabas (Jarrar and Hammouda, 2024) and WordNet (Miller et al., 1990). Later machine learning techniques such as Support Vector Machine (SVM) and Naive Bayes (Eid et al., 2010) became predominant, employing supervised learning techniques on labeled datasets. More recently, the rise of deep learning has significantly advanced the field, with neural network models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Al-Hajj and Jarrar, 2021b; Sun and Platoš, 2023). Transformer-based encoder models like BERT and its variants (Kenton and Toutanova, 2019) have further revolutionized WSD by leveraging large-scale pre-training on extensive corpora, followed by fine-tuning on specific WSD adapted datasets (Malaysha et al., 2023). Despite these advancements, Arabic NLP has seen slower progress in tackling WSD. To bridge this gap and stimulate further research, we present the ArabicNLU-2024 shared task. This initiative introduces a robust and rich datasets, aiming to propel Arabic NLU development and ultimately enhance human-computer interaction across diverse tasks.

Evaluation While this task is often better studied in English due to the availability of extensive resources, Arabic lacks adequate datasets and knowledge bases, necessitating the curation of high-quality resources to advance the research and support the Arabic NLP community (Elayeb, 2019). To address WSD, we have introduced the SALMA dataset (Jarrar et al., 2023c) as evaluation benchmark, which is a sense-annotated corpus with meanings extracted from two parallel lexicons: Al-Ghani Al-Zaher (Abul-Azm, 2014) and Contemporary Arabic Dictionary (Omar, 2008). This corpus comprises ~$34K$ tokens, all annotated with their candidate meanings, considering the relatedness of each sense to the actual meaning of the word in context. Although other corpora have been designed for Arabic WSD, none fully meet the task's requirements.

For instance, the Arabic version of the OntoNotes WSD dataset (Weischedel et al., 2013), annotated for three languages, lacks a well-defined sense list due to merging senses in a very coarse-grained manner. Similarly, the AQMAR dataset (Schneider et al., 2012) for Arabic WSD was not annotate using senses, but instead utilized high-level lexical classes. Nonetheless, we considered F1-score for evaluating the systems on SALMA because such metric represents balanced view for both precision and recall.

2.2 Location Mention Disambiguation (LMD)

Systems A few studies have addressed the LMD task for English language using machine learning and deep learning techniques. For example, Geoparspy (Middleton et al., 2018) uses SVM trained on gazetteer-based features. Additionally, Wang and Hu (2019) employed machine learning models for their toponym resolution system, including: (i) *DM_NLP* (Wang et al., 2019), a Light Gradient Boosting Machine (LightGBM), (ii) *UniMelb* (Li et al., 2019), an SVM classifier, and (iii) *UArizona* (Yadav et al., 2019), a heuristic-based system that favors toponyms with higher populations. Furthermore, Xu et al. (2019) proposed an attention-based model using two pairs of bi-LSTMs to match location mentions against the Foursquare gazetteer. The two-pair networks learn the left and right contexts of the LM, and both representations are processed through a fully connected layer for disambiguation.

Evaluation There is a dearth of public LMD datasets. In this shared task, we use the *only* public Arabic LMD dataset, IDRISI-DA (Suwaileh et al., 2023a), for evaluation. Discrete metrics such as Accuracy (Acc), Precision (P), Recall (R), and F_β scores are the most common metrics used to evaluate LMD systems (Zhang and Gelernter, 2014; Li et al., 2014; Ji et al., 2016; Middleton et al., 2018; Wang and Hu, 2019; Xu et al., 2019). However, these provide a broad overview and miss the nuances of different techniques. Distance based metrics assess LMD systems by measuring the great circle distance between the GPS coordinates of the gold and predicted location mentions, with overall performance computed by Median and Mean Error Distance. Acc, P, R, and F_β can also be computed within a distance d, commonly set to 161 km (100 miles). A significant issue with distance-based measures is the need to dynamically adjust the threshold for acceptable distance errors based on varying location granularity. While these measures are suitable for binary classification tasks, LMD is typically modeled as a multi-class classification or ranking task, making these measures less appropriate for evaluation. To address all these issues, we use Mean Reciprocal Rank at cutoff k (MRR@k).

2.3 Shared tasks

The ArabicNLU shared task is the first to address both word and location disambiguation in Arabic, marking a significant milestone in the field. This initiative is supported by other notable shared tasks aimed at understanding Modern Standard Arabic (MSA) and dialects. These include FinNLP for financial text processing (Malaysha et al., 2024) using the (Jarrar et al., 2023b) dataset, NADI for dialect identification (Abdul-Mageed et al., 2023) based on the (Abdul-Mageed et al., 2018) dataset, and WojoodNER for named entity recognition (Jarrar et al., 2024, 2023a) utilizing the Wojood dataset (Jarrar et al., 2022). Collectively, these collaborative efforts and interdisciplinary research projects foster a comprehensive understanding of linguistic nuances and enhance the applicability of NLP techniques across various contexts

3 Shared-task Overview

The ArabicNLU shared task consists of two primary sub-tasks, WSD and LMD. The WSD sub-task focuses on determining the correct semantic meaning of words (i.e., disambiguation of the word semantics) in a given context. The LMD sub-task, a special case of WSD, aims to accurately identify and disambiguate location mentions based on their geographical context.

The shared task mandates the use of pre-defined sense and location inventories that are directly linked to the provided datasets. Participants are prohibited from altering the senses, location mentions or toponyms within the test set. However, they are allowed to utilize external data and resources, including generative models, to improve their algorithms and models performance. To facilitate a unified evaluation, CodaLab,[1][2] a well-established platform for scoring shared task submissions, was employed. Furthermore, to guarantee equitable access to task guidelines and data, a dedicated web

[1] https://codalab.lisn.upsaclay.fr/competitions/17758
[2] https://codalab.lisn.upsaclay.fr/competitions/18918

page[3] was established for the shared task, providing detailed information to all participants.

We received registrations from 38 unique teams. During the testing phase, 4 teams submitted a total of 40 entries, among which 27 for the WSD subtask, and 13 for the LMD subtask. We received three description papers from the participated teams, all of which were accepted. Table 1 provides a detailed overview of the participated teams in alphabetical order by their name, including their affiliations and the tasks they participated in.

4 Subtask 1: Word Sense Disambiguation

4.1 Task Definition

Polysemous words that convey multiple meanings in different contexts have led to the emergence of the WSD task (Jarrar, 2021). WSD aims to determine the intended semantic meaning (i.e., sense) of a word within a given context (Al-Hajj and Jarrar, 2021a; Malaysha et al., 2023). Given a context c (i.e., a sentence), a target word w in c, and a set of candidate senses $S = \{s_1, ..., s_n\}$, for the target word w, the goal of the WSD task is to determine which of these senses is the intended meaning of w. Figure 2 depicts the WSD sub-task.

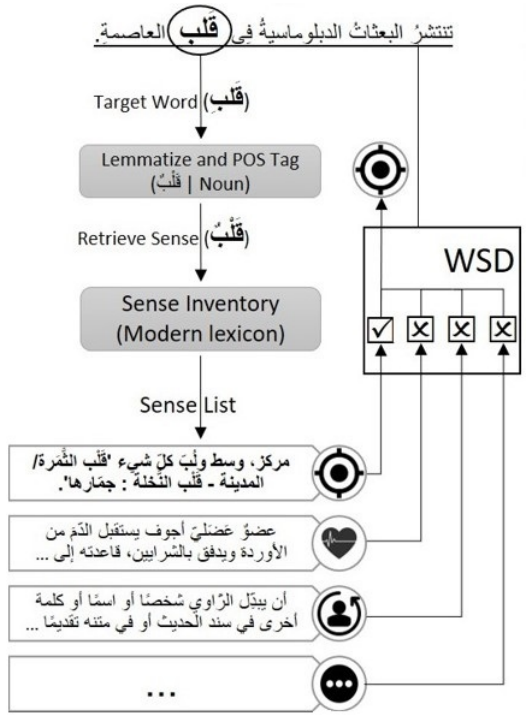

Figure 2: Illustration of the WSD subtask.

Participants were encouraged to utilize deep learning and generative methods for the WSD task. They were provided with a sense-annotated corpus of 1,340 sentences, part of the SALMA corpus (Jarrar et al., 2023c). Each target word in a given sentence has a set of candidate senses (glosses). Participants' submissions is expected to be in JSON format and must include the sentence id corresponding to the context c, the word id of w, and the sense id of the sense s_i (from the candidate senses S) for each target word in the test set.

4.2 Dataset: SALMA

SALMA corpus (Jarrar et al., 2023c), the first sense-annotated Arabic corpus, contains 1,440 sentences and 34k tokens, including 8,760 unique tokens and 3,875 unique lemmas. Manually annotated with 4,151 senses, it includes 19,030 nouns, 2,763 verbs, 7,116 functional words, and 5,344 punctuation marks and digits, as detailed in Table 2. The data was collected from 33 online media sources in Modern Standard Arabic (MSA), and has a 92% inter-annotator agreement (IAA) measured using Quadratic Weighted Kappa.

In the shared task, participants received development and test sets. The development set includes 100 sentences with corresponding candidate senses S and the correct sense s_i for each target word w in a given sentence c. The remaining 1,340 sentences were reserved for the test set, which included candidate senses, but excluded the correct sense. No training set was provided to encourage adoption and evaluation of generative model techniques, and participants were encouraged to use external datasets, sense inventories, or lexicons in their systems.

4.3 Baselines

Our WSD baseline approach involved developing a BERT-based system using Target Sense Verification (TSV) models. The TSV model is trained on a binary classification task that assigns confidence scores for True and False labels to each context-gloss pair. We created context-gloss pairs for each word in SALMA with varying context sizes to assess their impact on accuracy. The intended meaning was determined by ranking the glosses based on their True confidence scores, then selecting the one with the highest score as the intended gloss. Table 3 presents our baseline model' performance using Accuracy across diverse context window sizes. For instance, a window size of 11 encompassed five words on each side of the target word in the sur-

[3] https://sina.birzeit.edu/nlu_sharedtask2024/

Team	Affiliation	Task
Pirates (Wael et al., 2024)	Nile University	WSD
Rematchka (Abdel-Salam, 2024)	Cairo University	WSD, LMD
Upaya (Rajpoot et al., 2024)	SCB DataX	WSD, LMD

Table 1: Overview of participated teams and their tasks.

Term	Nouns	Verbs	Func. Words	Punct. & Digits
Total Tokens	19,030	2,763	7,116	5,344
Unique Tokens	6,670	1,593	322	175
Unique Lemmas	2,904	677	119	175
Unique Senses	3,151	792	206	2

Table 2: Statistics of SALMA corpus.

rounding context, while full context refers to the entire sentence. Our best-performing WSD baseline model achieved an accuracy (F1-score) of 84.2%.

Context Window Size	Baselines (F1-score)
3	82.80%
5	84.00%
7	83.80%
9	83.50%
11	**84.20%**
full	82.80%

Table 3: Baselines of WSD.

4.4 Participants' Systems

Thirty five teams registered for the WSD subtask, out of which only three teams submitted their system descriptions as shown in Table 1. Next we explore their approaches and results.

UPAYA (RAJPOOT ET AL., 2024): They leveraged LLMs, specifically Llama3 (AI@Meta, 2024) and GPT-4 (OpenAI, 2023), utilizing zero-shot learning techniques. The team employed a prompt-based approach where they manually crafted a natural language task description to be used consistently across all experiments. Initially, they experimented with a basic prompt that outputs plain text, then they enhanced the prompt by adding instructions that structures the input and output in JSON format to improve the model's comprehension. This structured format showed notable improvements with the Llama-3-70B-Instruct (AI@Meta, 2024) model. Additionally, they explored in-context learning by providing example sentences, target words and definitions in the prompt.

PIRATES (WAEL ET AL., 2024): The approach used by this team involves leveraging transformer-based models, specifically focusing on AraBERTv2 (Antoun et al., 2020), through three main experiments: using Sentence Transformers with Siamese networks (Ou et al., 2023), the SetFit framework (Pannerselvam et al., 2024), and a classification approach. The first experiment involves fine-tuning AraBERTv2 as a Sentence Transformer with contrastive loss, where the model learns to differentiate between positive and negative senses of a word within a sentence by calculating the Euclidean distance between their embeddings. This method uses a combined dataset prepared by integrating two resources, Al-Ghani Al-Zaher lexicon (Abul-Azm, 2014) and Arabic Context Gloss pairs (El-Razzaz et al., 2021), to ensure the model is exposed to both positive and hard negative samples. In the second experiment, they utilize the SetFit framework optimized for few-shot learning, which is advantageous due to its efficiency with minimal data input. This approach involves training the model on the sentence, target word, and its meaning, all separated by special tokens, and applying a cosine similarity loss function. The third experiment employs a more traditional classification approach using a transformer model for sequence classification. The AraBERTv2 model is fine-tuned with the SALMA development dataset, with the input structured similarly to the SetFit approach, but using the AdamW optimizer and training for fewer epochs. This method has shown the highest performance in terms of F_1-score among their three experiments.

REMATCHKA (ABDEL-SALAM, 2024): The participants employed zero-shot learning using LLMs and fine-tuning of pre-trained language models (PLMs). They explored the effectiveness of different models such as LLama3, WizardLM-2 (Xu et al., 2023), AceGPT (Huang et al., 2023), and OpenChat (Wang et al., 2023). In the zero-shot setting, the models were instructed to select the appropriate sense from a list of senses given the context and target word. This approach aimed to leverage the general language understanding capabilities of the models to perform WSD without task-

specific training. Additionally, fine-tuning models like MARBERT (Abdul-Mageed et al., 2021) and AraBERT (Antoun et al., 2020) were explored to enhance performance in WSD tasks.

4.5 Results

Table 4 summarizes the final results of the participating teams on the test dataset. The top-performing team, UPAYA, achieved a 78% accuracy with zero-shot learning technique using Llama3-70B-Instruct, outperforming GPT-4. In the second-place PIRATES, attained a 71% accuracy by fine-tuning a sense classifier using AraBERTv2, a model proven effective for Arabic. In the third place is REMATCHKA, which employed multiple generative models for zero-shot learning, but their prompt-based approach yielded the lowest performance with a 56%.

Notably, none of the participants surpassed our baseline (84.2%). This may suggest that generative models utilized by the participants, specially in zero-shot settings, still fall short of outperforming an encoder-based model fine-tuned on a discriminative task using high-quality large dataset. Generative models are also limited in their multilingual support as the majority of their training data covers English language. For instance, only 5% of the Llama3-70B training data is multilingual, covering 30 languages. It also an open question, whether the embeddings of causal generative models are less effective then bi-directional transformers for classification tasks.

Team	F_1**-score**
Baseline	84.2%
Upaya	77.8%
Pirates	70.8%
Rematchka	57.5%

Table 4: Results of participants on WSD subtask test data.

5 Subtask 2: Location Mention Disambiguation

5.1 Task Definition

LMD represents a challenging problem in retrieval and classification, primarily due to issues such as the lack of context, toponymic polysemy, and toponymic homonymy (Suwaileh et al., 2023a). Figure 3 presents a high-level overview of the task.

We formally define LMD problem as follows: Given a post p, the list of location mentions in p $L_p = \{l_i : i \in [1, n_p]\}$, where n_p is the number of location mentions in p, and a gazetteer $G = \{t_j : j \in [1, n_G]\}$, where n_G is the number of toponyms in G, an LMD system aims to match every location mention l_i in p to a toponym t_j in G that accurately represents it, if exists. Otherwise, the system must abstain and declare that l_i is unresolvable.

We perceive the LMD task as a candidate retrieval and ranking problem. For each location mention l_i, the LMD system must retrieve a ranked list of up to three candidate toponyms R from OpenStreetMap (OSM), where $R \subset G$. Toponyms retrieved by R are ranked based on the probability that each candidate is the correct toponym for l_i. Therefore, the LMD problem can be typically decomposed into two sub-problems: (i) candidate retrieval, which aims to retrieve a list of candidate toponyms from G, and (ii) candidate reranking, which aims to rerank the retrieved candidates R in order of likelihood.

5.2 Dataset: IDRISI-DA

The IDRISI-DA dataset was created in two phases: extracting location mentions (Suwaileh et al., 2023b) and disambiguating them (Suwaileh et al., 2023a). It is the first Arabic manually-labeled LMD dataset, designed with a particular attention on domain and geographical generalizability. Figure 4 shows the distribution of location types in IDRISI-DA, per disaster event, showing its domain and geographical coverage, therefore exhibiting a reasonable dialectical coverage (Suwaileh et al., 2023b). It includes 2,869 posts from X platform in diverse dialects, featuring 3,893 location mentions, with 763 unique mentions across seven countries. The dataset is split per event in ratios of 70:10:20 for training, development, and test sets, respectively. Each location mention in IDRISI-DA is annotated with only one correct toponym extracted from OSM, containing attributes such as geo-coordinates, location type, and addresses, among others.

5.3 Baseline

We compare the performance of the participated systems against OSM, a simple and common baseline for geolocation tasks. We specifically use Nominatim[4] that runs over the official OSM online gazetteer.[5]

[4] https://nominatim.org

[5] https://www.openstreetmap.org/

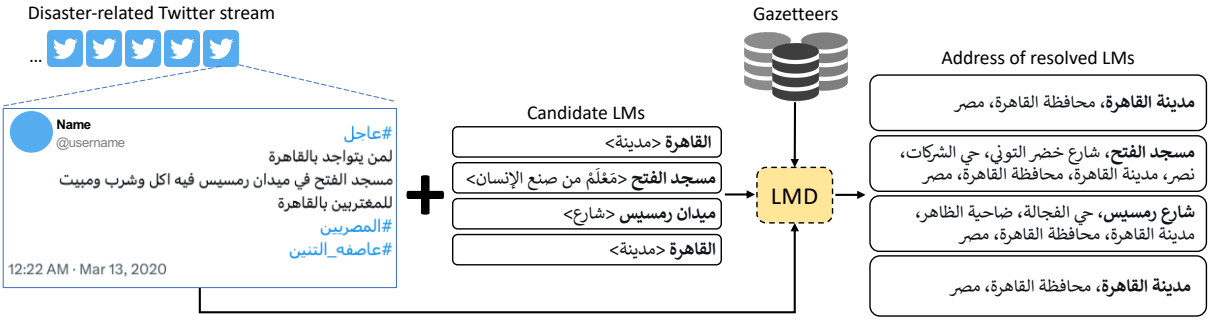

Figure 3: High-level overview of the LMD task.

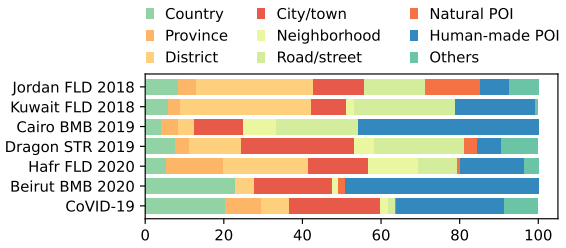

Figure 4: Distribution of location types in IDRISI-DA.

5.4 Participants' Systems

The LMD task had attracted 25 registered teams, however, only 2 of them managed to submit runs that we describe next.

REMATCHKA (ABDEL-SALAM, 2024): Used Llama3 (AI@Meta, 2024) to translate location mentions of type city or country to English if their type is verified. If not, Llama3 is queried for the most accurate country in English where the location mention is located, based on the post's context. The output is then passed to GeoPy[6] to retrieve corresponding toponyms. The rationale for translating to English is the GeoPy's degraded performance on Arabic text.

UPAYA (RAJPOOT ET AL., 2024): Proposed two retrieval stages approach. For every location mention, the system retrieves candidate toponyms from OSM, then re-ranks candidates using Cohere rerank-multilingual-v2.0.[7] The Cohere reranker involves self-attention mechanisms and transformer-based architectures that capture the similarity between location mentions and candidate toponyms.

5.5 Results

We present the *MRR@k* results of the participated systems in Table 5. The results demonstrate that both participants' systems outperform the baselines in *MRR@1*, highlighting their effectiveness in retrieving the correct toponym from OSM at the top rank. Notably, the REMATCHKA system substantially outperforms all other systems across all measures, exhibiting superior performance. This indicates the robustness of GeoPy, particularly when used with the English language. These results underscore the need for developing more robust models for Arabic LMD.

Team	MRR@1	MRR@2	MRR@3
$OSM_{baseline}$	0.5724	0.6396	0.6428
REMATCHKA	**0.9497**	**0.9500**	**0.9500**
UPAYA	0.5994	0.5994	0.5994

Table 5: Results of LMD participants on test set.

6 Conclusion and Future work

In this paper, we present the results of the ArabicNLU 2024 shared task, focusing on the challenges of Word Sense Disambiguation and Location Mention Disambiguation in Arabic Natural Language Understanding. The findings from participated teams highlight the ongoing challenges and research gaps associated with these subtasks. We observe that generative models underperform traditional classification architectures trained on labeled data. This is specially the case for low resourced languages which are not well supported in LLMs. LLMs are mostly English-centric due to the imbalanced training corpora. This also extends to other systems and tools such as GeoPy, requiring machine translation to English to achieve the desired performance. The challenge of multilingual support becomes even more apparent when working with Arabic dialectical data. To really democratize Arabic NLP, it is essential to compile large datasets in various Arabic dialects, as demonstrated by the work of (Jarrar et al., 2023d; Haff

[6] https://geopy.readthedocs.io
[7] https://cohere.com/blog/rerank

et al., 2022; Jarrar et al., 2023d). Additionally, new techniques must be developed to address the scarcity of dialectical data, and LLMs specifically tailored for the Arabic language need to be trained.

Our vision for this shared task is to create a collaborative environment that accelerates research and development in Arabic NLU. By facilitating the evaluation and comparison of various models and techniques, we aim to uncover new insights, foster innovation, and build a strong foundation of resources. This effort seeks to overcome current obstacles and significantly advance the capabilities of Arabic NLU technologies.

Limitations

Acknowledging the inherent constraints within the ArabicNLU shared task datasets is crucial. The SALMA dataset, utilized for the WSD subtask, primarily employs an extended version of "Modern" as a sense inventory, which includes referral glosses that need specific handling for broader applicability. Furthermore, it is limited to MSA, excluding dialects, and focuses only on single-word lemma senses.

The IDRISI-DA dataset, being crawled from X platform, faces significant limitation in its application due to recent X platform API restrictions that may reduce its utility for research focused on social media platforms. However, the dataset facilitates developing LMD systems the process informal text sourced from various platforms beyond X platform. Furthermore, fine-grained locations and temporary locations are underrepresented in IDRISI-DA, those are pivotal during emergencies. Nevertheless, the goal of this shared task is to develop generic LMD systems not domain(disaster)-specific ones.

Ethics Statement

The datasets provided for this shared task are derived from public sources, eliminating specific privacy concerns. The results of the shared task will be made publicly available to enable the research community to build upon them for the public good and peaceful purposes. Our data and techniques are strictly intended for non-malicious, peaceful, and non-military purposes.

Acknowledgements

This research is partially funded by the research committee at Birzeit University. We extend our gratitude to Taymaa Hammouda for the technical support. The contribution of Reem Suwaileh is partially funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund part of Qatar Research Development and Innovation Council (QRDI).

References

Reem Abdel-Salam. 2024. Rematchka at arabicnlu shared task: Evaluating large language models for arabicword sense and location sense disambiguation. In *The Second Arabic Natural Language Processing Conference (ArabicNLP 2024) Part of ACL 2024*.

Mohammed Alaeddine Abderrahim and Mohammed El Amine Abderrahim. 2022. Arabic word sense disambiguation for information retrieval. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 21(4):69:1–69:19.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7088–7105. Association for Computational Linguistics.

Miuru Abeysiriwardana and Deshan Sumanathilaka. 2024. A survey on lexical ambiguity detection and word sense disambiguation. *CoRR*, abs/2403.16129.

Abdul-Ghani Abul-Azm. 2014. Al-ghani al-zaher dictionary. *Rabat: Al-Ghani Publishing Institution*.

AI@Meta. 2024. Llama 3 model card.

Moustafa Al-Hajj and Mustafa Jarrar. 2021a. Arab-GlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Moustafa Al-Hajj and Mustafa Jarrar. 2021b. LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2vec Performance in Arabic Word-in-Context Disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 748–755, Online. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Vedanth Baiju. 2022. Word sense disambiguation in the domain of sentiment analysis through deep learning.

Wided Bakari, Mabrouka Ben-Sghaier, and Mahmoud Neji. 2021. Implementing an arabic question answering system using conceptual graphs. In *Hybrid Intelligent Systems - 21st International Conference on Hybrid Intelligent Systems (HIS 2021), December 14-16, 2021*, volume 420 of *Lecture Notes in Networks and Systems*, pages 295–304. Springer.

Sylvio Barbon Junior, Paolo Ceravolo, Sven Groppe, Mustafa Jarrar, Samira Maghool, Florence Sèdes, Soror Sahri, and Maurice Van Keulen. 2024. Are Large Language Models the New Interface for Data Pipelines? In *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments*, BiDEDE '24, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

M Soha Eid, Almoataz B Al-Said, Nayer M Wanas, Mohsen A Rashwan, and Nadia H Hegazy. 2010. Comparative study of rocchio classifier applied to supervised wsd using arabic lexical samples. In *Proceedings of the tenth conference of language engeneering (SEOLEC'2010), Cairo, Egypt*.

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A Maghraby. 2021. Arabic gloss wsd using bert. *Applied Sciences*, 11(6):2567.

Bilel Elayeb. 2019. Arabic word sense disambiguation: a review. *Artif. Intell. Rev.*, 52(4):2475–2532.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + Baladi: Towards a Levantine Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic. *CoRR*, abs/2309.12053.

Mustafa Jarrar. 2021. The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023a. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 748–758. ACL.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023b. ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 276–287. ACL.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Mustafa Jarrar and Tymaa Hasanain Hammouda. 2024. Qabas: An Open-Source Arabic Lexicographic Database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13363–13370, Torino, Italy. ELRA and ICCL.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023c. SALMA: Arabic Sense-annotated Corpus and WSD Benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023d. Lisan: Yemeni, Irqi, Libyan, and Sudanese Arabic Dialect Copora with Morphological Annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. 2016. Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 1271–1281, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2021. Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization. *Comput. Linguistics*, 47(4):813–859.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multimodal retrieval for visual word sense disambiguation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13053–13077. Association for Computational Linguistics.

Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. 2014. Effective location identification from microblogs. In *2014 IEEE 30th International Conference on Data Engineering*, pages 880–891, Chicago, IL, USA. Institute of Electrical and Electronics Engineers (IEEE).

Haonan Li, Minghan Wang, Timothy Baldwin, Martin Tomko, and Maria Vasardani. 2019. UniMelb at SemEval-2019 task 12: Multi-model combination for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1313–1318, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Nasser, Ismail Berrada, and Houda Bouamor. 2024. AraFinNLP 2024: The First Arabic Financial NLP Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Sanad Malaysha, Mustafa Jarrar, and Mohammed Khalilia. 2023. Context-Gloss Augmentation for Improving Arabic Target Sense Verification. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*. Global Wordnet Association.

Dastan Hussen Maulud, Subhi RM Zeebaree, Karwan Jacksi, Mohammed A Mohammed Sadeeq, and Karzan Hussein Sharif. 2021. State of art for semantic analysis of natural language processing. *Qubahan academic journal*, 1(2):21–28.

Stuart E. Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, 36(4):1–27.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Sangita S Modi and Sudhir B Jagtap. 2018. Web page classification using wsd and yago and ontology. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 887–891. IEEE.

Ahmed Mukhtar Omar. 2008. Contemporary arabic dictionary.(i1). *World of Books, Cairo, Egypt. Retrieval Date*, 14(8):2020.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Lizhen Ou, Yiping Yao, Xueshan Luo, Xinmeng Li, and Kai Chen. 2023. Contextad: Context-aware acronym disambiguation with siamese BERT network. *Int. J. Intell. Syst.*, 2023:1–14.

Kathiravan Pannerselvam, Saranya Rajiakodi, Sajeetha Thavareesan, Sathiyaraj Thangasamy, and Kishore Ponnusamy. 2024. Setfit: A robust approach for offensive content detection in tamil-english code-mixed conversations using sentence transfer fine-tuning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 35–42.

Gayatri Potey, Rucha Jadhav, Kushagra Shroff, Anish Gore, D Phalke, and J Shimpi. 2020. Fake news detection and sentiment analysis in twitter. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 8:72–75.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3668–3675. European Language Resources Association.

Pawan Kumar Rajpoot, Ashvini Kumar Jindal, and Ankur Parikh. 2024. Upaya at arabicnlu shared-task: Arabic lexical disambiguation using large language models. In *The Second Arabic Natural Language Processing Conference (ArabicNLP 2024) Part of ACL 2024*.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: An arabic case study. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short*

Papers, pages 253–258. The Association for Computer Linguistics.

Yujia Sun and Jan Platoš. 2023. Attention-based stacked bidirectional long short-term memory model for word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023a. IDRISI-D: arabic and english datasets and benchmarks for location mention disambiguation over disaster microblogs. In *Proceedings of ArabicNLP 2023, Singapore (Hybrid), December 7, 2023*, pages 158–169. Association for Computational Linguistics.

Reem Suwaileh, Muhammad Imran, and Tamer Elsayed. 2023b. IDRISI-RA: The first Arabic location mention recognition dataset of disaster tweets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16298–16317, Toronto, Canada. Association for Computational Linguistics.

Tasneem Wael, Eman Elrefai, Mohamed Makram, Sahar Selim, and Ghada Khoriba. 2024. Pirates at arabicnlu2024: Enhancing arabic word sense disambiguation using transformer-based approaches. In *The Second Arabic Natural Language Processing Conference (ArabicNLP 2024) Part of ACL 2024*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *CoRR*, abs/2309.11235.

Jimin Wang and Yingjie Hu. 2019. Are we there yet? evaluating state-of-the-art neural network based geoparsers using eupeg as a benchmarking platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities '19, New York, NY, USA. Association for Computing Machinery.

Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM_NLP at SemEval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 917–923, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. *Philadelphia: Linguistic Data Consortium*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Canwen Xu, Jiaxin Pei, Jing Li, Chenliang Li, Xiangyang Luo, and Donghong Ji. 2019. Dlocrl: A deep learning pipeline for fine-grained location recognition and linking in tweets. In *The World Wide Web Conference, WWW 2019*, pages 3391–3397, San Francisco, CA, USA. ACM.

Vikas Yadav, Egoitz Laparra, Ti-Tai Wang, Mihai Surdeanu, and Steven Bethard. 2019. University of Arizona at SemEval-2019 task 12: Deep-affix named entity recognition of geolocation entities. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1319–1323, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70.