rematchka at ArabicNLU2024: Evaluating Large Language Models for Arabic Word Sense and Location Sense Disambiguation

Reem Abdel-Salam
Cairo University, Faculty of Engineering, Computer Engineering / Giza, Egypt
reem.abdelsalam13@gmail.com

Abstract

Natural Language Understanding (NLU) plays a vital role in Natural Language Processing (NLP) by facilitating semantic interactions. Arabic, with its diverse morphology, poses a challenge as it allows multiple interpretations of words, leading to potential misunderstandings and errors in NLP applications. In this paper, we present our approach for tackling Arabic NLU shared tasks for word sense disambiguation (WSD) and location mention disambiguation (LMD). Various approaches have been investigated from zero-shot inference of large language models (LLMs) to fine-tuning of pre-trained language models (PLMs). The best approach achieved 57% on WSD task ranking third place, while for the LMD task, our best systems achieved 94% MRR@1 ranking first place.

1 Introduction

Natural Language Understanding (NLU) is a crucial field in language processing that aims to comprehend and interpret human language. Word Sense Disambiguation (WSD) plays a crucial role in resolving multiple meanings and nuances of words, but applying WSD to the Arabic language presents unique challenges due to its rich morphology, dialectal variations across different regions, and polysemy. A special case within WSD is Location Mention Disambiguation (LMD), which involves disambiguating location references in Arabic texts considering diverse geographical, historical, and cultural contexts.

Developing effective NLU techniques for Arabic, particularly for WSD and LMD tasks, is of great significance for various applications such as information retrieval, machine translation, and natural language understanding systems. However, there are not enough good linguistic resources and datasets specifically designed for these tasks in Arabic, making it difficult to train and test models, affecting their accuracy and reliability.

This paper outlines our solution to the ArabicNLU Shared-Task (Khalilia et al., 2024), which involves two tasks. The first task is Word Sense Disambiguation (WSD), where the objective is to determine the correct meaning of a target word within a given context by selecting from a set of candidate senses (i.e., glosses or definitions) associated with the target word. The second task is Location Mention Disambiguation (LMD). In LMD, the goal is to match each location mention within a post or tweet to the correct toponym from a geo-positioning database that contains a list of toponyms. This matching process aims to find the toponym in the database that accurately represents the mentioned location.

To address the WSD challenge, we conducted experiments with large language models (LLMs) in a zero-shot setting. In this approach, we have instructed the model to select the appropriate sense from a list of senses given the context, and target word. Additionally, we explored the effectiveness of fine-tuning models like MARBERT and AraBERT for improved performance in WSD.

For the LMD task, we have experimented with LLMs in a zero-shot setting and in the context of the Retrieval-Augmented Generation (RAG) framework. In the Zero-shot setting, we provided instructions to the model to translate the location and provide the corresponding country name as a means to obtain the accurate toponym from the Geopy framework. In the RAG framework, we maintained a record of the context, location, country, and the valid toponyms that were available. Retrieving the toponym involved evaluating the similarity score between the new context and location, and comparing it with the knowledge base to select the highest-scoring match. The experiments aimed to determine which approach yielded better results and provided insights into the effectiveness of LLMs in solving WSD and LMD challenges in Arabic natural language understanding.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 383–392
August 16, 2024 ©2024 Association for Computational Linguistics

The rest of the paper goes as follows: section 3 gives an overview of the dataset, section 4 discusses the proposed methods, section 5 shows experimental results, and section 6 concludes the paper.

2 Related Work

This section of the paper provides a literature review on Arabic word sense disambiguation and location mention disambiguation. It discusses the importance of these tasks in accurate and context-aware natural language processing in Arabic. The review focuses on existing approaches, resources, and evaluation methodologies employed in the field. (Malaysha et al., 2023) proposed ArabGlossBERT dataset, which consists of 167K context-gloss pairs collected from Arabic dictionaries, is commonly used but relatively small. The authors present an enriched version of the dataset achieved through machine back-translation, increasing its size to 352K pairs. They evaluate the impact of augmentation using various data configurations to fine-tune BERT for the target sense verification (TSV) task. The accuracy ranges from 78% to 84% across different data configurations, with some improvements observed for specific parts-of-speech (POS) in certain experiments. (Al-Hajj and Jarrar, 2021) focuses on fine-tuning BERT models for Arabic WSD. The authors constructed a dataset of labeled Arabic context-gloss pairs, extracted from the Arabic Ontology and a lexicographic database. Each pair was labeled as true or false, and target words in each context were identified and annotated. Three pre-trained Arabic BERT models were then fine-tuned using this dataset. The authors also experimented with different supervised signals to emphasize target words in context. The experiments yielded promising results, achieving an accuracy of 84%, even when dealing with a large set of senses in the experiment. (Kaddoura and Nassar, 2024) introduces a new dataset for Arabic WSD and proposes a novel approach using BERT. The proposed WSD system outperforms existing methods, achieving an approximate F1-score of 96%. The effectiveness of WSD is demonstrated in a case study involving sentiment analysis as a downstream task. (Saidi et al., 2023) addresses the challenges of WSD in Arabic due to limited resources and semantic sparsity. The authors propose WSDTN, a manually annotated corpus consisting of 27,530 sentences collected from diverse sources. Each sentence includes a target word and its corresponding sense. They also present a transformer-based model for disambiguating new words and evaluate the corpus's performance. The baseline approach achieves an accuracy of approximately 90%. (Saidi and Jarray, 2023) presents a hybrid approach for Arabic Word Sense Disambiguation (AWSD) by combining a single-layer Convolutional Neural Network (CNN) with contextual representation using BERT. The proposed approach utilizes a concatenation of BERT models as word embeddings to capture both target and context representations. Experimental results demonstrate that the proposed model surpasses state-of-the-art approaches, achieving an accuracy of 96.42% on the Arabic WordNet dataset, thereby improving the performance of WSD in Arabic languages.

3 Data

In subtask 1, the development dataset provided by the organizers is a subset of the SALMA corpus (Jarrar et al., 2023). The training dataset consisted of 100 sentences, while the test set comprised 1,340 sentences. As for subtask 2, the training dataset used in this task was derived from the IDRISI-D dataset (Suwaileh et al., 2023), which contained 2,170 sentences. The validation and test sets were composed of 333 and 791 sentences, respectively. Within the training dataset, there were a total of 3,893 location mentions, with 763 unique mentions.

4 System Overview

LLMs have proven to excel in a wide range of linguistic tasks, due to their extensive training on vast amounts of internet data. With their high capabilities in tasks such as generation and reasoning. LLMs have extensively been studied and utilized in languages like English. However, their potential in the Arabic language has not been fully explored, presenting a significant opportunity to investigate and harness LLMs for Arabic NLU tasks. To seize this opportunity, we conduct experiments on several large language models that have been trained on data of multiple languages including Arabic, namely LLama3, WizardLM-2 (Xu et al., 2023), AceGPT (Huang et al., 2023), and openchat (Wang et al., 2023), to assess their performance in word sense disambiguation and location mention disambiguation tasks. In addition, we compared their performance with pre-trained language models (PLM) such as MARBERT (Abdul-Mageed et al., 2021) and AraBERT (Antoun et al.). In the following section, we provide detailed information about their training and inference setup.

4.1 Subtask 1

To address this task, we explored two main approaches: the zero-shot setting with large language models (LLMs) and fine-tuning using pre-trained language models (PLMs). Extensive evidence has shown that by finetuning LLMs to adhere to natural language instructions, their performance can be significantly enhanced on tasks that were previously unseen, particularly in zero-shot and few-shot settings (Li et al., 2023; Liu et al., 2023). In our study on WSD task, we employed zero-shot settings for LLama3, WizardLM2, LLama3-Instruct, and AceGPT-7B. Different instructions and pre-processing techniques have been explored to improve LLMs performance. Refer to the Appendix A for the prompt/instructions and architecture. Moreover, previous studies have indicated that pre-trained language models (PLM) can outperform LLMs in certain specific tasks (Liusie et al., 2024; Kwon et al., 2023; Almazrouei et al., 2023). Building upon this insight, we decided to fine-tune two PLMs, namely MARBERT and AraBERT, in three distinct settings to evaluate their performance in word sense disambiguation. The first setting utilizes the dataset provided by the organizers of the shared task. In this setting, the development dataset is divided into training and dev subsets using an 80%-20% split. The second setting incorporates the dataset mentioned in the work by (Kaddoura and Nassar, 2024), along with the dev-dataset provided by the shared task organizers. In this setting, the validation set is 20% of both datasets. However, the validation set is carefully curated to ensure uniqueness based on words and senses that have not been encountered in the training split. The third setting involves the combination of three datasets: 1- provided by organizers, 2- in (Kaddoura and Nassar, 2024) work, 3- Razzaz [1]. In this setting, the validation set constitutes 20% of the entire combined dataset. The training for MARBERT and AraBERT goes as follows, the problem is formulated as a binary classification problem, for a given sentence, the word and sense/meaning model has to predict 1 if it is the correct meaning or zero if it is not. The input format is defined as follows: a window size of 10 words surrounding the target word, followed by a separator, the target word itself, another separator, and finally, the sense being conditioned on for prediction. During inference as shown in Figure 1, we select the sense/meaning that gives a high probability score to be true.

4.2 Subtask 2

In this task, the geo-positioning database contains a collection of location data with associated coordinates or geographic information. The toponyms refer to the names or labels assigned to these specific locations within the database. The task involves retrieving and ranking these location toponyms accurately based on the content or context provided in a given post. The goal is to identify the most relevant and appropriate location names that align with the information mentioned in the post. The architecture in this subtask is inspired by retrieval augmented systems (RAG). We first build a database from the training dataset composed of text provided, location mentioned in Arabic and English, country location in Arabic and English, and toponyms. To build this database, at first, we query LLama3 whether the given word is most likely a city or country name. If it's a country name we ask for its English equivalent and save it in the database. If it is not a country name, we query LLama3 to give the name of the most likely country, where this location would be in the given correct context (post). Then we ask LLama3 to translate both location and country names in English. The reason behind it is that GeoPy (tool) that is responsible for retrieving toponyms given names of cities or locations is more sensitive to the English language than Arabic. For instance, it could not retrieve the location of السفارة الكندية لبنان however, was able to retrieve it when given “Canadian Embassy, Lebanon". Another example toponyms retired by Geopy for location حفرالباطن doesn’t match any of toponyms in the training set for this location, however, when presented with its English equivalent “al-Hafar al-Batin, Saudi Arabia" the tool was able to retrieve all corresponding toponyms correctly. During inference, as shown in Figure 2, at first for each location mentioned, we determine whether the location is a country name or a city name, or a location in a country by querying LLama3. If it's a city/location name we ask LLama3 to give the most likely country that this location should be in. Then we match with what is in the database based on context, country, and location relevance. The cosine similarity

[1] data available at https://github.com/MElrazzaz/Arabic-word-sense-disambiguation-bench-mark

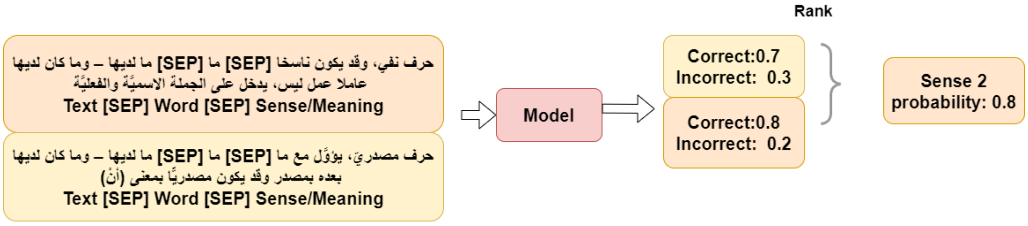

Rank
حرف نفي، وقد يكون ناسخا [SEP] ما [SEP] ما لديها – وما كان لديها
عاملا عمل ليس، يدخل على الجملة الاسميَّة والفعليَّة
Text [SEP] Word [SEP] Sense/Meaning
حرف مصدريّ، يؤوَّل مع ما [SEP] ما [SEP] ما لديها – وما كان لديها
بعده بمصدر وقد يكون مصدريًّا بمعنى (أنْ)
Text [SEP] Word [SEP] Sense/Meaning
Model
Correct:0.7
Incorrect: 0.3
Correct:0.8
Incorrect: 0.2
Sense 2
probability: 0.8

Figure 1: Inference pipeline.

score is calculated between the target query, location, and information in the database. If there is relevance (similarity score higher than 0.8), then we return back toponyms stored in the database. If there is no relevance, Geopy is queried based on the English version, if there is no response, then it is queried by the Arabic name. The query contains the location and country name. Refer to the Appendix B for the prompt/instructions and architecture.

5 Results and Discussion

This section presents the model's performance during the development and testing phases using the official metrics. Additionally, an error analysis is carried out to pinpoint the shortcomings of the presented models. For subtask 1 the official metric is the macro average F1-score, while for subtask 2 the official metric is the mean reciprocal rank (MRR@k), where k indicates the cutoff of the retrieved ranked list of candidate toponyms. For dev-phase results refer to Appendix C.

5.1 Test-phase results

Table 1 shows the results of developed models for subtask1. LLama3-Instruct model with a limited context window of ±10 words around target words showed the best performance with 57.52% accuracy. While comes in second place LLama3 with a limited context window of ±10 words around target words, with 57.34% accuracy. As shown from the results, limited context windows matter when inference with LLMS. However, size of context window is a hyperparameter that needs to be adjusted. PLMs struggle in performance for WSD tasks. Most of the models struggle to outperform the baseline by a large margin. The proposed approach archives third place in the leaderboard with a 20% margin from the first-place solution. Table 2 shows the results of developed models for subtask2. Refer to Appendix B for details about pre-processing and post-processing. The proposed approach outperforms the baseline by a margin of 37% and lands in first place. In addition, it outperforms the second place solution by a margin of 35%.

Model	Approach	Accuracy
LLama3	No limit to context length	54.14
LLama3 Instruct	Limit length of context ±10 words around target word	57.52
LLama3	Limit length of context ±10 words around target word	57.34
LLama3	Limit length of context ±5 words around target word	56.09
LLama3 Instruct	Limit length of context text ±5 words around target word	55.18
LLama3	Instruction in English	55.63
MARBERT	Training Setting 1	37.7
MARBERT	Training Setting 2	38.32
MARBERT	Training Setting 3	39.34
Baseline	Context Window ±11	84.20

Table 1: Performance of the submitted models on the leaderboard in subtask 1.

5.2 Error analysis

In this section, we will explore the limitations and challenges of the model, with the goal of identifying areas for improvement and enhancing its overall performance and reliability. Specifically, for subtask 1, we encountered the following issues:

- LLMs struggled to infer word meaning from

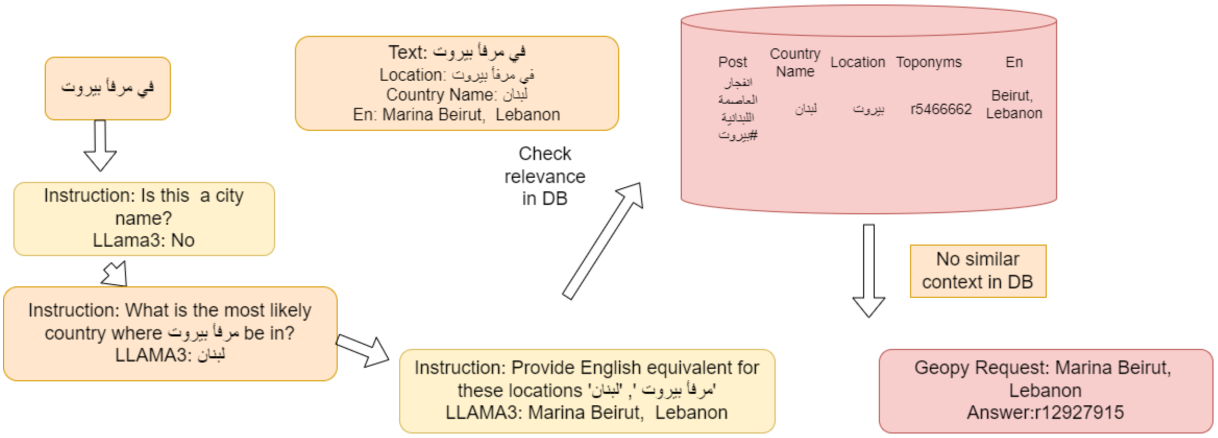

في مرفا بيروت
Instruction: Is this a city name?
LLama3: No
Instruction: What is the most likely country where مرفأ بيروت be in?
LLAMA3: لبنان
Text: في مرفأ بيروت
Location: في مرفأ بيروت
Country Name: لبنان
En: Marina Beirut, Lebanon
Check relevance in DB
Post
Country Name
Location
Toponyms
En
انفجار العاصمة اللبنانية #بيروت
لبنان
بيروت
r5466662
Beirut, Lebanon
No similar context in DB
Instruction: Provide English equivalent for these locations 'لبنان' ,' مرفأ بيروت'
LLAMA3: Marina Beirut, Lebanon
Geopy Request: Marina Beirut, Lebanon
Answer:r12927915

Figure 2: Inference pipeline.

Model	Approach	Accuracy
LLama3	Zero-shot No pre-processing or No post-processing	30.06
LLama3	Zero-shot Pre-processing	30.74
LLama3	Data Base	92.2
LLama3	Data Base post-processing	94.97
Baseline	-	62.70

Table 2: Performance of the submitted models on the leaderboard in subtask 2.

the context window, especially when dealing with large texts. This resulted in the model often responding with a generic message, such as "How could we help you?"

- LLMs did not consistently respond in the requested format, making it difficult to extract answers and leading to random sense selection at times.
- Instructions were provided to the model to determine the correct number of senses based on context and a sense choice list, but some instances presented the correct sense as text instead of a numerical value.
- Some senses provided to the model included verses number from the Quran as an example the meaning of the word آحرف مصدريّ ظرفيّ آوَأَوْصَانِي بِالصَّلَاةِ وَالزَّكَاةِ مَا دُمْتُ حَيًّا مريم ٣١ : مدّة دوامي حيًّا , which posed challenge. As the model answer the number of the verse rather than the choice number.
- Some senses were too large for the model to handle effectively, causing difficulties in responding to the original task.
- Certain models struggled to understand the task and continued providing context instead of solving the task itself.
- PLMs faced issues due to the limited number of training datasets and a small context window of 512 tokens, making it challenging to correctly infer meaning for large senses, which was more common in the test set than in the training set.

For subtask 2, one of the most common errors was the LLMs (LLama3) inability to correctly associate locations with countries, for instance it would associate النيل with Iraq rather than Egypt. Another problem is the translation of location from Arabic to English for instance 1/3 of the time it would translate مرفأ بيروت ، لبنان to “Market in Beirut, Lebanon". Refer to the Appendix D for more analysis.

6 Conclusion

This paper presented the work conducted during the ArabicNLU Shared Task. The best solution for both subtasks was based on LLama3, which demonstrated superior performance compared to MARBERT and AraBERT. However, it is important to note a limitation observed during the experiments, namely the issue of hallucination, where the LLama3 model occasionally provided incorrect output formats, or would continue text rather than solving task. Future work will focus on enhancing these models to address this limitation and further improve their accuracy and reliability

References

Muhammad Abdul-Mageed, AbdelRahim 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 (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. Arabglossbert: 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.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, et al. 2023. Alghafa evaluation benchmark for arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023. SALMA: Arabic sense-annotated corpus and WSD benchmarks. In *Proceedings of ArabicNLP 2023*, pages 359–369, Singapore (Hybrid). Association for Computational Linguistics.

Sanaa Kaddoura and Reem Nassar. 2024. Enhancedbert: A feature-rich ensemble model for arabic word sense disambiguation with statistical analysis and optimized data collection. *Journal of King Saud University-Computer and Information Sciences*, 36(1):101911.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond english: Evaluating llms for arabic grammatical error correction. *arXiv preprint arXiv:2312.08400*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Yilun Liu, Shimin Tao, Xiaofeng Zhao, Ming Zhu, Wenbing Ma, Junhao Zhu, Chang Su, Yutai Hou, Miao Zhang, Min Zhang, et al. 2023. Automatic instruction optimization for open-source llm instruction tuning. *arXiv preprint arXiv:2311.13246*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

Sanad Malaysha, Mustafa Jarrar, and Mohammed Khalilia. 2023. Context-gloss augmentation for improving Arabic target sense verification. In *Proceedings of the 12th Global Wordnet Conference*, pages 254–262, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Rakia Saidi and Fethi Jarray. 2023. Stacking of bert and cnn models for arabic word sense disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(11):1–14.

Rakia Saidi, Fethi Jarray, Asma Akacha, and Wissem Aribi. 2023. Wsdtn a novel dataset for arabic word sense disambiguation. In *International Conference on Computational Collective Intelligence*, pages 203–212. Springer.

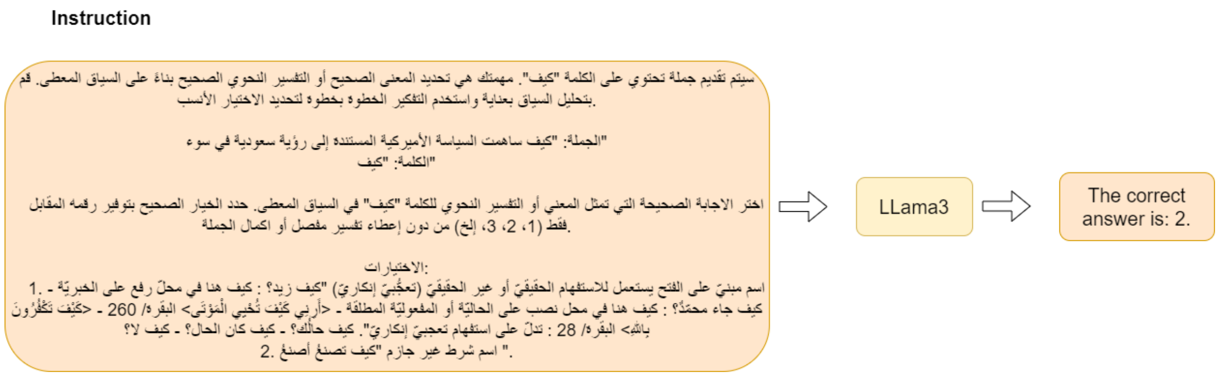
Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023. Idrisi-d: Arabic and english datasets and benchmarks for location mention disambiguation over disaster microblogs. In *Proceedings of ArabicNLP 2023*, pages 158–169.

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

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. *arXiv preprint arXiv:2304.12244*.

A Subtask-1 Methodology

To address the first subtask, the LLMs were instructed to select the most suitable definition. The input was the target word, context, and list of sense choices. Various instructions have been investigated to improve model performance. The figure 3 shows the whole pipeline.



سيتم تقديم جملة تحتوي على الكلمة "كيف". مهمتك هي تحديد المعنى الصحيح أو التفسير النحوي الصحيح بناءً على السياق المعطى. قم بتحليل السياق بعناية واستخدم التفكير الخطوة بخطوة لتحديد الاختيار الأنسب.

الجملة: "كيف ساهمت السياسة الأميركية المستنده إلى رؤية سعودية في سوء"
الكلمة: "كيف"

اختر الاجابة الصحيحة التي تمثل المعني أو التفسير النحوي للكلمة "كيف" في السياق المعطى. حدد الخيار الصحيح بتوفير رقمه المقابل فقط (1، 2، 3، إلخ) من دون إعطاء تفسير مفصل أو اكمال الجملة.

الاختيارات:
1. اسم مبنيّ على الفتح يستعمل للاستفهام الحقيقيّ أو غير الحقيقيّ (تعجُّبيّ إنكاريّ) "كيف زيد؟ : كيف هنا في محلّ رفع على الخبريّة - كيف جاء محمّدٌ؟ : كيف هنا في محل نصب على الحاليّة أو المفعوليّة المطلقة - <أَرِنِي كَيْفَ تُحْيِي الْمَوْتَى> البقره/ 260 - <كَيْفَ تَكْفُرُونَ بِاللهِ> البقره/ 28 : تدلّ على استفهام تعجبيّ إنكاريّ". كيف حالُك؟ - كيف كان الحال؟ - كيف لا؟
2. اسم شرط غير جازم "كيف تصنعُ أصنعُ ".

Figure 3: Inference pipeline.

A.1 Instruction Designs

The following set of instructions has been investigated, Including both the Arabic version of the instruction as well as the English version

1. اختر الاجابة الصحيحة التي تمثل المعني أو التفسير النحوي للكلمة المعطاة في الجملة التالية

   الجملة : []

   الكلمة : []

   الاختيارات: اختر المعنى الصحيح لكلمة [] من خلال تقديم الرقم المقابل (١، ٢، ٣، الخ) فقط دون شرح.

   Choose the correct answer that represents the meaning or grammatical interpretation of the word given in the following sentence

   Sentence: []

   Word: []

   Choices: Choose the correct meaning of the word [] by providing only the corresponding number (1, 2, 3, etc.) without explanation.

2. سيتم تقديم جملة تحتوي على الكلمة []. مهمتك هي تحديد المعنى الصحيح أو التفسير النحوي الصحيح بناءً على السياق المعطى. قم بتحليل السياق بعناية واستخدم التفكير الخطوة بخطوة لتحديد الاختيار الأنسب.

   الجملة : [] الكلمة : []

   اختر الاجابة الصحيحة التي تمثل المعني أو التفسير النحوي للكلمة [] في السياق المعطى. حدد الخيار الصحيح بتوفير رقمه المقابل فقط (١، ٢، ٣، إلخ) من دون إعطاء تفسير مفصل أو اكمال الجملة.

   الاختيارات: []

   A sentence containing the word [] will be presented. Your task is to determine the correct meaning or grammatical interpretation based on the given context. Carefully analyze the context and use step-by-step reasoning to determine the most appropriate choice.

   Sentence: [] Word: []

   Choose the correct answer that represents the meaning or grammatical interpretation of the word [] in the given context. Select the correct option by providing only its corresponding number (1, 2, 3, etc.) without giving a detailed explanation or completing the sentence.

   Choices: []

3. Instructions:
   You will be presented with a text containing the word "[]". Your task is to determine the correct meaning or grammatical pattern based on the given context. Carefully analyze the context and use step-by-step reasoning to select the most appropriate choice.

   Text: "[]"
   Word: "[]"

   Which of the following choices best represents the meaning or grammatical pattern of the word "" in the given context? Choose the correct option based on careful analysis.

> Choices:
>
> "[]"
>
> Please provide the number corresponding to the correct choice.

A.2 Pre-processing

To handle large contexts and inconsistencies in answers, various pre-processing techniques have been used. At first, we remove any punctuation or numbers from sentences. Then we limit input sentence/context/text to be ±5 or ±10 words around the target word. Post-processing is applied to get the answer, if the model answers in numeric values we extract it. Otherwise, the model could answer with textual values of the number of another number included in the sense, in this case, we don't handle it. But if the model returned the whole sentences of the chosen sense, we match it with the current sense list and return the number of the sense.

A.3 Hyperparameters

Hyperparameter	Value
Learning-rate	4e-5 or 5e-4
Schedular	cosine-annealing
Weight decay	1e-2
Epochs	30
Optimizer	AdamW
Metric	F1-macro on dev-set

Table 3: The full hyperparameter search space.

The hyperparameter used to to fine-tune PLMs is listed in Table 3.

B Subtask-2

B.1 Instruction

The following set of instructions was used to get information from LLama3. For detecting whether a given word is a country or location we use

“According to this text: []

Does this word [] refer to a country name

The answer should be yes/no."

For translating words we use the following instruction:

“Translate this word to English:[]. The answer must not contain any Explanation and should be structured as "Answer: " " " For querying location we use the following instruction

“ Given this context: "[]"

According to the previous text, Which country is the location "[]" located in?

If you could not get it from context, which country is highly likely to have this location/place "[]" in?

The answer should be in the English language, with no explanation needed only, and should be structured as "Answer: Country ".

"

B.2 Pre-processing and Post-processing

In this approach, the impact of pre-processing and post-processing on results has been explored. In the pre-processing step, both tweet/text and location mention are clean, that is to remove hashtags, mentions, emojis, and underscores. In post-processing, various steps are applied

1. If Geopy returns nothing for the correct text, we try to remove one of the words in the text. For example, Geopy could not retrieve any information about مدرسة العذارية عجلتون ، لبنان however, by removing the last word مدرسة العذارية ، لبنان it was able to retrieve the correct information.

2. Another step in applying filtering based on retrieved toponyms, is if country and location are not explicitly present we query LLama3 whether both the location and country are the most likely to be in toponyms. If it answered no then we remove it else we keep it.

3. For ranking, if the country name or location name is explicitly in toponyms then it has high ranking than those its not in it.

C Dev-phase results

Table 4 shows the performance of various LLMs models in zero-shot setting inference on the training dataset in subtask 1. LLama showed the best performance with around 70% accuracy compared to other models. While WizardLM2 comes in second place. Table 5 shows the performance of MARBERT and AraBERT during dev-phase in subtask 1 on various settings. Both Models showed comparable results and improvement when having a large dataset during training. In subtask 2, LLama3-Instruct with Database, pre and post-processing steps achieved 90% F1-score.

Model	Approach	Accuracy
LLama3	English Prompt	54.453
LLama3	Limit length of context ±10	58.097
LLama3	No Context limit	57.89
Openchat	English prompt	57.489
Openchat	Limit length of context ±10	49.030
Openchat	No Context limit	48.78
AceGPT	Limit length of context ±10	50.862
WizardLM2	Limit length of context ±10	52.777
WizardLM2	Limit length of context ±5	56.126

Table 4: Performance of the LLMs models on the training set for subtask 1.

Model	Approach	Accuracy
MARBERT	Training Setting 1	86
MARBERT	Training Setting 2	75
MARBERT	Training Setting 3	77
AraBERT	Training Setting 1	50
AraBERT	Training Setting 2	75
AraBERT	Training Setting 3	76

Table 5: Performance of the PLMs models on the dev-set for subtask 1.

D Error Anaylsis

Extensive error analysis has been conducted on LLMs for both subtasks.

D.1 Subtask-1

The following section will discuss errors in each model.

D.1.1 AceGPT-7B

AceGPT-7B fails most of the time to solve the original problem, for instance, for the following instruction

سيتم تقديم جملة تحتوي على الكلمة آكيفْ. مهمتك هي تحديد المعنى الصحيح أو التفسير النحوي الصحيح بناءً على السياق المعطى. قم بتحليل السياق بعناية واستخدم التفكير الخطوة بخطوة لتحديد الاختيار الأنسب.

الجملة: آكيف ساهمت السياسة الأميركية المستندة إلى رؤية سعودية في سوءْ الكلمة: آكيفْ

اختر الاجابة الصحيحة التي تمثل المعني أو التفسير النحوي للكلمة آكيفْ في السياق المعطى. حدد الخيار الصحيح بتوفير رقمه المقابل فقط (١، ٢، ٣، إلخ) من دون إعطاء تفسير مفصل أو اكمال الجملة.

الاختيارات: ١. اسم مبنيّ على الفتح يستعمل للاستفهام الحقيقيّ أو غير الحقيقيّ (تعجُّبيّ إنكاريّ) آكيف زيد؟ : كيف هنا في محلّ رفع على الخبريّة - كيف جاء محمّدٌ؟ : كيف هنا في محل نصب على الحاليّة أو المفعوليّة المطلقة - البقرة ٢٦٠ - البقرة ٢٨ : تدلّ على استفهام تعجبيّ إنكاريٌّ. كيف حالُك؟ - كيف كان الحال؟ - كيف لا؟

٢. اسم شرط غير جازم آكيف تصنعُ أصنعُ آ.

it would return

كيف ساهمت السياسة الأميركية المستندة إلى رؤية سعودية في سوء ٣. مبتدأ إنكاري آكيف لاْ، وهو القسم الثالث من الجملة الشرطيةْ.

٤. العبارة التعجبية آكيف ساهمت ...؟آ وهي الاستخدام الأكثر قوة لـ آكيفْ، حيث تُظهر الدهشة والسؤال.

٥. جزء من السؤال الاعتراضي - كيف؟

٦. اسم خبر أو جملة شرطية ءذا كيفْ، حيث آكيفْ هي الصيغة الاعتراضية لـ ‹‹إذاْ.

٧. بداية الجملة السؤالية - كيفيًا، ومن الأمثلة: آكيف تُحى الْمَوْتَىْ وْكِيفَ تُصرَفُ زكَاةِكْ.

٨. خبر إذا، عندما، لو كنت ست - كيف؟

D.1.2 Opengpt

For openGPT it fails when Quran verses are presented in text. It would either return them all or answer “ **These are some of the meanings and**

contexts in which "" can be used. Remember that Arabic is a rich language with many nuances, so it's essential to consider the context when using words to ensure the intended meaning is conveyed. "

D.1.3 WizardLM2

The failure cases occurred because the model attempted to provide additional information and simplify the text while elaborating on the chosen sense. As a result, it became challenging to automatically determine the correct answer. For instance it would return مجموعة محلية لها تاريخ محلي ومصالح محلية . However choices are اسم منسوب إلى مَحَلّ and داخليّ، متعلِّق بموضع معيّن أو خاصّ بمنطقة

D.1.4 LLama3

The failure cases observed in LLama3 share similarities with WizardLM2. Both models tend to simplify the content, and in some instances, LLama3 mistakenly selects the number mentioned inside the choice text instead of the actual number corresponding to the choice itself. These challenges highlight the need for further improvements in the models' understanding of context and accurate interpretation of the provided choices.

D.2 Subtask-2

LLama3 struggled in correctly associating locations with countries and correctly translating the name of the location. For instance 1/3 of the time it would translate مرفأ بيروت to “Market in Beirut" or “Marina Beirut". Another example المنيل would translate it to “The key" or “The Treasure". Therefore we ask the model 10 times the same question and take the most repetitive answer. Another problem is associating location to country for instance location المنيل or معهد الاورام would be associated 1/3 time with Lebanon, 1/3 of the time with Iraq, and 1/3 of the time with Egypt. However, this was solved by adding “most probable country and is usually a known place " in the prompt. Another Example location رأس تنورة 1/3 percent of the time it will associated with “'Sudan', 1/3 of the time with “Lebanon" and 1/3 of the time with “Saudi Arabia". Another problem is in Geopy itself, incorrectly retrieving toponyms itself. For instance, when given السفارة الكندية لنان, it would retrieve a lot of places not related to the current query itself. Another problem with Geopy, is that if a location is presented in multiple countries it would return them all, even if the query does contain the correct country name.