Greeting
We would like to extend an invitation to participate in the KSAA-CAD: Contemporary Arabic Dictionary Shared Task! For Reverse Dictionary and Word Sense Disambiguation at ArabicNLP 2024.
Please find all the necessary information below.
https://arai.ksaa.gov.sa/sharedTask2024/

Registration deadline: 29th of April 2024.

Registration:

Participants need to register via this link:
https://docs.google.com/forms/d/e/1FAIpQLSetmIT7XQ3megI6dGI9xm17wc-oY86z_UAxunVoRCBsbFK7Lg/viewform !

Tasks:

Task 1: Reverse Dictionary (RD)

RDs, identified by their sequence-to-vector format, characterized as sequence-to-vector, introduce a differentiated strategy in contrast to traditional dictionary lookup methods. The RD task concentrates on the conversion of human-readable glosses into word embedding vectors.
This process entails reconstructing the word embedding vector corresponding to the defined word, a methodology aligning with the approaches of (Mickus et al., 2022; Zanzotto et al., 2010; Hill et al., 2016).
The dataset includes, lemma, lemma vector representations, and their respective gloss.
The developed model is expected to generate novel lemma vector representations for the unseen human-readable definitions in the test set. Such a strategy allows users to search for words based on anticipated definitions or meanings.

Task 2: Word Sense Disambiguation (WSD)

WSD focuses on identifying the specific sense of a word in a given context. This technique involves determining a word's intended meaning by calculating the overlap between its contextual use and the provided gloss or definition.
The dataset consists of word, context, and context ID, and corresponding gloss ID. The developed model is expected to retrieve the suitable gloss ID for the word in the context from the WSD dictionary.

Datasets:

Datasets can be downloaded from the CODALAB platform:

- CODALAB-Task1 RD: https://codalab.lisn.upsaclay.fr/competitions/18510
- CODALAB-Task2 WSD: https://codalab.lisn.upsaclay.fr/competitions/18468

This section details the structure of the JSON dataset files provided.

Task 1: RD:

The dataset itself comprises two core components: the dictionary data and the word embedding vectors.

In the first iteration of KSAA-RD (Al-Matham et al., 2023), the dataset derived from a single source: the "Contemporary Arabic Language Dictionary" by Ahmed Mokhtar Omar (Omar, 2008). In this revised edition, we endeavor to expand our sources to encompass three dictionaries of Contemporary Arabic Language. The first of these is the "Contemporary Arabic Language Dictionary" by Ahmed Mokhtar Omar (Omar, 2008), a resource previously utilized in the first iteration KSAA-RD. The second is the newly released dictionary of the Arabic contemporary language "Mu'jam Arriyadh" (Altamimi et al., 2023). The third is the "Al Wassit LMF Arabic Dictionary" (Namly, 2015).

These dictionaries comprise words, commonly referred to as lemmas, and these may come with glosses, part of speech (POS), and examples.
In the generation of these word embeddings, our approach is to utilize three distinct architectures of contextualized word embedding, such as Electra (Clark et al., 2020) and BERT (Devlin et al., 2019), to enhance the effectiveness of the system. Specifically, we employ AraELECTRA (Antoun et al., 2021), AraBERTv2 (Antoun et al., 2020), and camelBERT-MSA (Inoue et al., 2021) —referred to respectively as electra, bertseg, and bertmsa—for our methodologies.
As a concrete instance, here is an example from the training dataset for the Arabic dictionary:

```
{
"id":"ar.45",
"word":"عين",
"gloss":"عضو الإبصار في ...",
"pos":"n",
"electra":[0.4, 0.3, …],
"bertseg":[0.7, 2.9, …],
"bertmsa":[0.8, 1.4, …],
 }
```

Task 2: WSD

The dataset itself comprises two core components: the WSD context gloss mapping and dictionary data. The WSD context gloss mapping consists of word, context, context ID, and corresponding gloss ID. As a concrete instance, here is an example from the training dataset for the corresponding WSD JSON:

```
{
"context_id":"context.301"
"context":"يأتي برمجان اللغة العربية...",
"word": "اللغة",
"gloss_id":"gloss.305"
"lemma_id":"ar.200"
}
```

The dictionary data contains word, gloss, and gloss ID. The dictionary data is derived from the "Contemporary Arabic Language Dictionary" by Ahmed Mokhtar Omar (Omar, 2008). As a concrete instance, here is an example from the WSD dictionary:

```
{
"lemma_id": "ar.200",
"gloss_id":"gloss.305",
"gloss":" كُلُّ وسيلة لتبادل المشاعر والأفكار كالإشارات ..."
}
```

Task	Train	Dev	Test

RD Entries	31,372	3,921	3,921
WSD Entries	22,404	2,801	2,801
WSD dictionary	15,865		

Baselines :

The baseline repository for both tasks is available here.

Task 1: RD:

We leverage SOTA MARBERT (Abdul-Mageed, 2021) and CamelBERT-MSA (Inoue et al., 2021) models, employing fine-tuning techniques to excel in Arabic RD. These models are SOTA, proven by their superior performance in the shared task of KSAA-RD (Al-Matham et al., 2023), representing a winning approach.

Task 2: WSD:

Initially, the dataset is enriched using lemma id by joining WSD entries with WSD dictionary to incorporate both relevant and irrelevant glosses, and cleaned. The model is trained to determine the relevance of a gloss to a word in context. The highest probability gloss is then calculated for each word in context, improving its ability to accurately identify context-appropriate meanings. We employ two approaches for WSD:

- Fine-tuning: The approach leverages BertForSequenceClassification, specifically with CamelBERT-MSA and AraBERTv2, due to their exceptional precision in identifying context-sensitive words. The target word in context is wrapped with special tokens "<token>word</token>".
- Neural Network: This approach involves feeding the three text embeddings (context, word, and gloss) from the multilingual-E5-base model into a simple LSTM neural network consisting of a 3D input layer, a single LSTM layer, a dense layer, and an output layer. The integrating these advanced embeddings enhances the LSTM's ability to accurately distinguish and disambiguate word senses, lead to improve performance. We refer to this configuration as the E5+LSTM model.

Important dates:

- Release of training, dev data, and evaluation scripts: 15th of March 2024.
- Release of test data: 15th of April 2024.
- Registration deadline: 29th of April 2024.
- End of the evaluation cycle (test set submission closes): 3th of May 2024.
- Results released: 29th of April 2024
- System description paper submissions due: 10st of May 2024.

Contact:

- Email: waad.wtss@gmail.com
- Email: r.almatham@gmail.com

Awards:

We are pleased to announce the awards for the Arabic Reverse Dictionary Shared Task at ArabicNLP 2024. The top-ranked teams in each task will receive cash prizes as follows:

Task 1: Reverse Dictionary (RD)
1st Ranked: $350
2nd Ranked: $250
3rd Ranked: $150

Task 2: Word Sense Disambiguation (WSD)
1st Ranked: $350
2nd Ranked: $250
3rd Ranked: $150

Best regards!!