Rosetta Stone at the Arabic Reverse Dictionary Shared Task: A Hop From Language Modeling To Word–Definition Alignment

Ahmed ElBakry *
Microsoft, Egypt
ahmedlebakry@microsoft.com

Mohamed Gabr
Microsoft, Egypt
mohamed.gabr@microsoft.com

Muhammad ElNokrashy
Microsoft, Egypt
muhammad.elnokrashy@microsoft.com

Badr AlKhamissi *
EPFL, Switzerland
badr.alkhamissi@epfl.ch

Background

Reverse dictionaries: Systems to find a word based on its description. They tackle the **tip-of-the-tongue** phenomenon [1].

Method	Learned Parameters	Features	Characteristics
Heuristic	Possible	Strings	Match Description vs. Dictionary Definition
Neural	Yes	Vectors	Match Description vs. Dictionary Definition Embeddings
Handcrafted	Possible	Inspired by human intuition	Match by human-inspired features between Description and Dictionary Definition
Pretrained LMs	Yes	Contextual Vectors	Match by contextual embeddings of Description and Dictionary Definition

Results

✨ Interesting!
Average of **CamelBERT-MSA** & **MARBERTv2** output embeddings

Subtask	Embedding	MSE	Cosine	Rank	P@1	P@10
Subtask 1	SGNS	**.152**/.161	.645/.637	.242/.214	.031/.034	.099/.114
	ELECTRA	**.030**/.035	.605/.552	.254/.281	.445/.414	.597/.540
Subtask 2	SGNS	**.170**/.180	.659/.624	.127/.204	.185/.120	.407/.355
	ELECTRA	**.053**/.048	.400/.387	.320/.372	.312/.316	.375/.389

Table 4: Results on TestSet / DevSet for both subtasks. MSE is Mean-Squared-Error. P is Precision.

Datasets

1. The Arabic Language Dictionary
- **ELECTRA** embedding
- **SGNS** embedding
- Gloss (definition of the word)
- POS tag

2. The English Language Dictionary
- Has 4,355 datapoints in total.
- Arabic and English glosses, IDs, and words.
- Arabic embeddings.

	Train	Dev	Test
Ar Dict	45,200	6,400	6,410
Ar-EN Map	2,843	299	1,213
Ar Dict	50,877	12,719	N/A

Table 1: Statistics about Data Sizes

	Example Word	Example Definition
Task 1	تحنن عليه	ترحم، تعاطف عليه و رحمه
Task 2	زور الكلام	To knowingly and willfully make a false statement of witness while in court

Table 2: Word-Definition Pairs Illustrating Subtasks 1 and 2.

Discussion

- **Exploring Cross-Lingual Alignment Further**
  - Training an autoencoder to enable cross-alignment between the Arabic and English dictionaries. Further exploration of this technique may yield promising results.
- **Augmenting Training Data Through Self Synthesis**
  - Finetuning AraT5 to jointly predict the embeddings from the encoder side, and predicting the definition of the word from the decoder side as extra supervision.

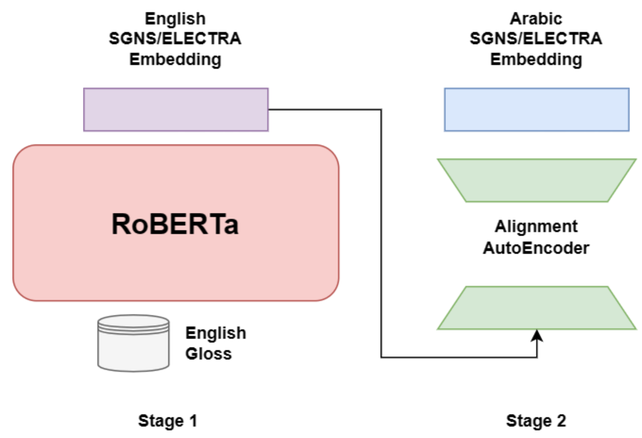

Figure 1: Method Explored for Subtask 2

System

1. Subtask 1: Arabic Definitions to Arabic Embeddings

- Finetuning x2 (SGNS & ELECTRA)
  - AraBERTv2
  - MARBERTv2
  - CamelBERT-MSA
  - CamelBERT-Mix
- For each Base LM from the list, we **finetune** model to **minimize MSE** between **word and definition**.
- The final model is an average ensemble of CamelBERT-MSA and MARBERTv2.

Hyperparameter	Value
Batch Size	100
lr	1.0e-4
lr Scheduler	OneCycleLR
pct	0.2
Weight Decay	1.0e-4
Epochs	20
Optimizer	AdamW

Table 3: Hyperparameters Used

2. Subtask 2: English Definitions to Arabic Embeddings

1. Same as [3], we use the Arabic translation of the English definitions as input to our finetuned Arabic models.
2. This approach enables the reuse of models and solutions that were initially developed for subtask 1.

Conclusion

- **Solving the Arabic Reverse Dictionary Shared Task:** Select an Arabic word given an Arabic or English definition.
- **Arabic Pretrained LMs as Base Models:** AraBERTv2, MARBERTv2, CamelBERT-MSA, CamelBERT-Mix
- **Finetuning and Ensembling:** Improve capture of semantic and syntactic aspects of input definitions and corrects for small errors.
- **First Task:** We predict a projection of model output to minimize MSE to Electra and SGNS embeddings of word and definition.
- **Second Task:** The solution translates the English definitions into Arabic and reuses the models from the first subtask.

References

- **[1]** Roger Brown and David McNeill. 1966. The “tip of the tongue” phenomenon.
- **[2]** Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT:Deepbidirectional transformers for Ara bic.
- **[3]** Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, An gela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. ArXiv, abs/2305.14240