AREEj: Arabic Relation Extraction with Evidence

Osama Rakan Al Mraikhat and **Hadi Hamoud** and **Fadi A. Zaraket**
Arab Center for Research and Policy Studies, Doha
{oalmraikhat, hhamoud, fzaraket}@dohainstitute.edu.qa

Abstract

Relational entity extraction is key in building knowledge graphs. A relational entity has a source, a tail and a type. In this paper, we consider Arabic text and introduce evidence enrichment which intuitively informs models for better predictions. Relational evidence is an expression in the text that explains how sources and targets relate. This paper augments the existing SRED^{FM} relational extraction dataset with evidence annotation to its 2.9-million Arabic relations. We leverage the augmented dataset to build AREEj, a relation extraction with evidence model from Arabic documents. The evidence augmentation model we constructed to complete the dataset achieved .82 F1-score (.93 precision, .73 recall). The target AREEj outperformed SOTA mREBEL with .72 F1-score (.78 precision, .66 recall).

1 Introduction

We define *relational extraction with evidence* (REE) as the task of extracting related entities (source and target), identifying the relation type between them, and *providing evidence from text to support the relations*. Relation Extracting (RE) is an important task to build knowledge graphs. Enriching knowledge graph edges with evidence labels helps in several ways: (i) it supports explainable AI tasks potentially by large language models (Pruthi et al., 2020), and (ii) it improves the performance of the RE models as evidence plays the role of a hint for an existing relation.

Evidence is important as it helps understanding Machine Learning (ML) decisions (Pruthi et al., 2020) and provides decision-makers with more confidence in ML model predictions. In turn, detecting evidence boosts the performance of ML models as it is a step in a chain-of-thought to infer the relation and its type (Wei et al., 2022).

A relational entity with evidence $ree = \langle s, t, s_{ne}, t_{ne}, rt, e \rangle$ where s is the source, t is the target, s_{ne} and r_{ne} are named entity types of the source and the target, rt is the relation type, and e is the evidence. The elements s, t, and e are extracted from the text itself, and the rest are predefined categories and classes. One document may contain multiple REEs and REEs may have elements in common.

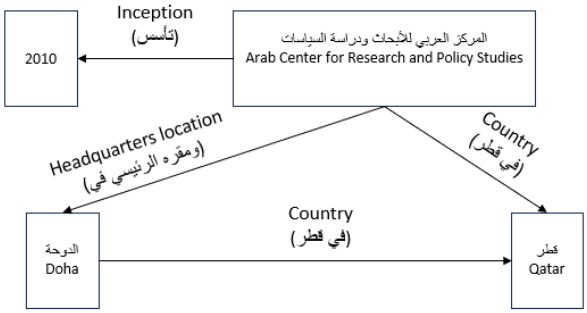

Figure 1: Knowledge graph with 4 extracted REEs. Evidence shows within parenthesis.

Figure 1 shows a sentence with four relations where the main entity s_1 entity is connected as source with three other entities: t_1 inception date (2010), t_2 headquarters location (Doha), and t_3 country (Qatar). Doha is also a source $s_2 = t_2$ connected to t_3 indicating it is in Qatar. The evidence shows between brackets explaining why the the model reported these relations and their types.

Researchers developed systems capable of extracting relations from Arabic text including a morphology-based and regular expression based approach (Jaber and Zaraket, 2017), and a cross-language learning approach (Taghizadeh et al., 2018) that leveraged knowledge transfer for the lack of large Arabic relational datasets.

Research leveraged Wikipedia data to build a dataset with 19 relation types (Zakria et al., 2019). Multilingual data from Wikipedia was also utilized to construct a larger dataset, across 14 languages and with 36 relation types (Seganti et al., 2021), where at most one relation was assigned per

Proceedings of The Second Arabic Natural Language Processing Conference, pages 67–72
August 16, 2024 ©2024 Association for Computational Linguistics

text. BERT (Devlin et al., 2019) was used to train two models:

- A sequence classification that classifies the type of the relation in the text, and
- An entity extraction that identifies the source and target entity of the relation.

The claim is that multilingual models achieve better results than single language models. The Arabic subset of the data contains 9,000 rows with 9 unique relation types.

Later work presents SRED^{FM} with 40-Million relations in 18 different languages to be the largest RE dataset up to our knowledge. Its Arabic subset covers 393 out of a total of 400 relation types with 2.9-Million relations. SRED^{FM} is extracted from Wikipedia using the cRocoDiLe tool (Huguet Cabot and Navigli, 2021). The mREBEL model was trained on SRED^{FM} to perform relational extraction (Huguet Cabot et al., 2023).

This paper presents work that augments the Arabic subset of SRED^{FM} with relational evidence, and then to use the resulting dataset to train AREEj, a relational extraction with evidence model. We perform the augmentation via fine-tuning an open-source large language model to extract evidence from the Arabic text given a relational sequence. The evidence augmentation task performed well with .82 F1-score, .93 precision and .73 recall AREEj extracted relation with evidence with .72 F1-score, .78 precision and .66 recall. We make the annotated dataset available online for the research community.

2 Related Work

RE started as a cascaded task where it first classifies source and target entities, and then identifies relation types between them. Recent research approached RE as a sequence-to-sequence (Seq2Seq) transformation. The raw text constitutes the input sequence, and the relation constitutes the output sequence.

REBEL extracts relational entities from English text excluding evidence (Huguet Cabot and Navigli, 2021). It is a seq2seq Bidirectional and Auto-Regressive Transformers (BART) model (Lewis et al., 2020) trained on linearized relations extracted automatically from Wikipedia using cRocoDiLe.

The mREBEL tool (Huguet Cabot et al., 2023) is trained and tested on the (SRED^{FM}) and (RED^{FM}) multilingual RE datasets extracted from Wikipedia data. Each relation is specified as a tuple wit source, target, entity types for source and target, and relation type. The training dataset was automatically generated, while the testing dataset was manually revised. mBART (Liu et al., 2020), which is typically used for translation, was used to map and normalize the relational types across all included 18 languages.

The work in (El Khbir et al., 2022) extracts named entities, relation types, and *event triggers* jointly. It covers 7 named entity types, 6 relation types, and 8 event types. The data is encoded by concatenating the last and third last BERT embeddings and extracting entities and events to identify the spans.

A rule-based model extracts the source, target, and relation spans depending on Part-of-Speech tags (Saber et al., 2022). The Stanford Arabic Parser (Green and Manning, 2010) and WordNet (Miller, 1994) were leveraged to extract triplets, which works well for sentences with simple structures.

In (Huang et al., 2021), an approach was introduced to extract relations from documents and use attention to extract evidence. Documents are multiple sentences and the evidence is several sentences that support a detected relation. This is different from our target evidence as we target fine-grained relations at the sentence level. We also extract the relation and the evidence simultaneously.

The work in (Ma et al., 2023) improves on memory efficiency and lack of annotations. Our work solves the annotations problem by using ChatGPT-4 to annotate the data with evidence, fine-tune an open-source LLM using the generated data, and then use the fine-tuned LLM to annotate 2.9 million relations with evidence.

To do that we leverage the chain of thought (CoT) technique in our prompts (Wei et al., 2022) where requiring intermediate steps in inference improves the quality of the final result and avoids inaccuracies and hallucinations.

The work in (Wadhwa et al., 2023) compares several LLM approaches for RE. This includes GPT-3 few-shot learning, a fine-tuned Flan-T5, and improved versions of both with CoT. The results recommend utilizing LLMs with CoT for RE tasks.

Task/Relation type	AREEj			mREBEL		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Extraction	.78	**.66**	**.72**	.89	.56	.69
Classification	.9	**.69**	**.78**	.92	.58	.71
Manufacturer	**1.00**	**.9**	**.95**	.82	.82	.82
Country	**.73**	.96	.83	.72	1.00	.84
Country of citizenship	1.00	1.00	1.00	1.00	1.00	1.00
Director	1.00	**1.00**	**1.00**	1.00	.75	.86
Inception	.86	1.00	.92	1.00	1.00	1.00

Table 1: AREEj and mREBEL results for extraction, classification, and selected relation types.

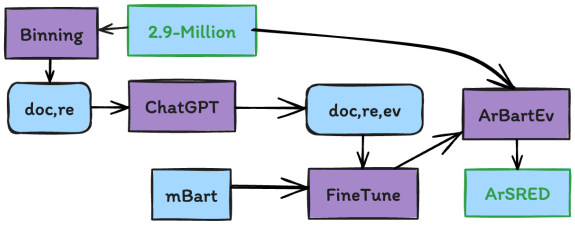

Figure 2: Building ArSRED flow diagram

3 Relational Data with Evidence

We constructed ArSRED via augmenting the 2.9-Million relations Arabic subset of SRED^{FM} with evidence as illustrated in Figure 2. We did this in several steps.

Construct representative documents and relations subset (RDRS). We selected a subset of Arabic SRED^{FM} to be augmented with seed evidence annotations. Intuitively, we targeted extraction of relations from smaller text to boost the GPT context performance and limit the number of tokens used later on with ChatGPT-4. We selected the documents with length between five and 80 words inclusive. This happened to be around 89% of the 2.9 million documents.

Then we split the set based on the relation type and the source and target entity types resulting in 2,526 bins that represent all typology combinations present in the set. Then we selected a representative sample (20 relations) from each bin starting from the smaller bins. We ended up with 68,183 relations selected from their corresponding 13,180 documents. Note that the total number of relations is higher than $20 * 2,526$ as when we include a document, we also keep all its relations and not only the ones from the selecting bin.

Construct Base Evidence Annotations. We used ChatGPT-4 to find evidence for the RDRS relations. We optimized ChatGPT-4 performance with a CoT based prompt which improved its performance.

We tried several prompts with a variety of examples. The final prompt consisted of two examples and four relations. We provided an explanation of the intermediate steps to get the evidence for each relation. We passed only one document per explanation. The output is the text of the evidence or "none".

Documents with plenty of relations resulted in less evidence elements returned. So we added the specific number of elements to the prompt. This partially fixed the problem. This augmented 55,078 RDRS relations with evidence and the remaining 13,105 relations with "none' forming the base evidence annotations.

We repeated the same process with the whole Arabic subset of the testing RED^{FM} and humanly revised the results for 345 documents, and 864 relations with 32 relation types. We consider this as our testing set.

Fine-tune LLM for evidence annotations. We used the base evidence annotations to fine-tune ArBartEv from the mBART (Liu et al., 2020) model. ArBartEv transforms an Arabic input text with a relation tuple, into an evidence of the relation in a sequence-to-sequence manner.

ArSRED Annotation with evidence. We passed the whole 499,638 documents with 2.9 million relations Arabic subset of SRED^{FM} to ArBartEv. to obtain ArSRED set augmented with evidence. A small 6% of the predictions came with a corrupted output sequence structure. We noticed upon inspection that these are mostly large documents, and we marked the evidence as "none".

The following is an example of how ArBartEv takes a sentence with a relation and produces a relation with evidence to construct the training dataset.

Input:

هذا الهرم، المعروف أيضاً باسم الهرم الأحمر الشمالي، هو أكبر هرم يقع في دهشور في الجيزة في مصر…

s: الأحمر الهرم ,location; t: دهشور ,location;rt:location

Output:

s: الأحمر الهرم ,location; t: دهشور ,location;rt:location; e: يقع في

4 Training ArBartEv and AREEj

ArBartEv requires relations to produce evidence. ArBartEv takes $x = \langle t, r \rangle$ composed of text t and relation r and produces $y = \langle r, e \rangle$ where e is the evidence.

$$p(y = r, e|x) = \prod_{i=1}^{\text{len}(y)} p_{\text{mBART}}(y_i|y_{<i}, x) \quad (1)$$

This produced dataset ArSRED as discussed in Section 3. The intuition behind including r in both input and output is that transformers tend to work better when asked to perform tasks in steps and with hints.

We trained AREEj based on the mBART seed model and ArSRED to take an Arabic text t and produce a relation r with evidence e.

$$p(y = r, e|t) = \prod_{i=1}^{\text{len}(y)} p_{\text{mBART}}(y_i|y_{<i}, t) \quad (2)$$

We target only Arabic text so we limited the tokenizer to Arabic and kept the named and relation entity types as special tokens as specified in REBEL and mREBEL.

Model fine-tuning was executed on 1x NVIDIA RTX A6000 48 GB VRAM GPU. The hyperparameters used are .00005 learning rate, 5000 warm-up steps, 16 batch size, and 156140 max steps. The rest are the default hyperparameters of mBART.

5 Results and Evaluation

For evidence equivalence for ArBartEv, we computed the number of word-based intersections between its evidence results and those produced by ChatGPT-4 on RED$^{\text{FM}}$. For entries with no intersection, we count that as a miss. For entries with intersections, we performed a manual check and a human expert deemed the extracted evidence as true or not. ArBartEv achieved an F1-score of .82 on evidence extraction (P=.93,R=.73).

For AREEj, we performed a beam search at the output level and obtained six predictions. Note that these predictions may have relations that share relational elements. We noticed that some extracted relations were correct as evaluated by a human expert, yet the testing set missed them. AREEj correctly identifies 72 relations that mREBEL misses, while mREBEL identifies 47 relations AREEj misses. AREEj achieves .72 F1-score (p=.78, r=.66) with correct evidence .86% of the time as detailed in the Extarction row of Table 1. AREEj was trained on 2.9-Million relations with evidence annotations while mREBEL was trained on 40-Million relations in 18 languages without evidence annotations. Table 1 also shows that AREEj outperforms mREBEL in relation classification which measures whether any relation (regardless of type) exists between two detected entities.

Fourteen relation types are detected by AREEj but not by mREBEL. Examples include 'employer', 'use', 'member of', 'industry', 'main subject', and 'date of death'. Four relation types are detected by mREBEL and not by AREEj including 'characters', and 'member of sports team'. Lower rows of Table 1 compare AREEj and mREBEL on selected relation types from ArSRED.

6 Conclusion

Adding evidence annotations resulted in extracting more relations and relation types. Recall increased significantly and precision suffered slightly. In future work, we plan to use the extracted evidence to improve precision via relating to relation and entity types and reducing the "none" samples to improve recall.

7 Limitations

The testing dataset contained only 32 relation types. This is an initial limitation of RED$^{\text{FM}}$. Relations in text exist that are not reported in RED$^{\text{FM}}$

requiring manual checks. Augmenting RED^{FM} with instances covering the 393 relations should be considered in future work.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Niama El Khbir, Nadi Tomeh, and Thierry Charnois. 2022. ArabIE: Joint entity, relation and event extraction for Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 331–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Spence Green and Christopher D. Manning. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 394–402, Beijing, China. Coling 2010 Organizing Committee.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315, Online. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. Red^{fm}: a filtered and multilingual relation extraction dataset. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Ameen Jaber and Fadi A. Zaraket. 2017. Morphology-based entity and relational entity extraction framework for arabic. *ArXiv*, abs/1709.05700.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3965–3970, Online. Association for Computational Linguistics.

Yasser Mohamed Saber, Hala Abdel-Galil, and Mohamed Abd El-Fatah Belal. 2022. Arabic ontology extraction model from unstructured text. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6066–6076.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

Nasrin Taghizadeh, Heshaam Faili, and Jalal Maleki. 2018. Cross-language learning for arabic relation extraction. *Procedia Computer Science*, 142:190–197.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Gehad Zakria, Mamdouh Farouk, Khaled Fathy, and Malak Makar. 2019. Relation extraction from arabic wikipedia. *Indian Journal of Science and Technology*, 12:01–06.