Cross-Lingual Knowledge Projection and Knowledge Enhancement for Zero-Shot Question Answering in Low-Resource Languages

Sello Ralethe and **Jan Buys**
Department of Computer Science, University of Cape Town, South Africa
rltsel002@myuct.ac.za, jbuys@cs.uct.ac.za

Abstract

Knowledge bases (KBs) in low-resource languages (LRLs) are often incomplete, posing a challenge for developing effective question answering systems over KBs in those languages. On the other hand, the size of training corpora for LRL language models is also limited, restricting the ability to do zero-shot question answering using multilingual language models. To address these issues, we propose a two-fold approach. First, we introduce LeNS-Align, a novel cross-lingual mapping technique which improves the quality of word alignments extracted from parallel English-LRL text by combining lexical alignment, named entity recognition, and semantic alignment. LeNS-Align is applied to perform cross-lingual projection of KB triples. Second, we leverage the projected KBs to enhance multilingual language models' question answering capabilities by augmenting the models with Graph Neural Networks embedding the projected knowledge. We apply our approach to map triples from two existing English KBs, ConceptNet and DBpedia, to create comprehensive LRL knowledge bases for four low-resource South African languages. Evaluation on three translated test sets show that our approach improves zero-shot question answering accuracy by up to 17% compared to baselines without KB access. The results highlight how our approach contributes to bridging the knowledge gap for low-resource languages by expanding knowledge coverage and question answering capabilities.

1 Introduction

Knowledge bases (KBs) like ConceptNet (Speer et al., 2016), Freebase (Bollacker et al., 2007) and DBpedia (Mendes et al., 2012) represent factual information as knowledge triples, expressing relations between concepts or real-world entities. KBs are constructed either by automatic extraction from monolingual corpora (Mendes et al., 2012; Bollacker et al., 2007) or through contributions by target language speakers (Mitchell et al., 2018; Speer et al., 2016). Knowledge bases are important for NLP applications such as question answering and information retrieval.

KB construction for low-resource languages faces challenges due to insufficient availability of Wikipedia pages or other relevant monolingual data in the target language. This often results in their omission from multilingual KBs or incomplete coverage when included. These limitations prevent low-resource languages from benefiting from recent NLP advancements such as entity alignment for knowledge graphs (Chen et al., 2017; Zhang et al., 2019; Mao et al., 2020; Zhu et al., 2021). The scarcity of comprehensive KBs in low-resource languages limits their ability to leverage NLP applications such as machine translation (Moussallem et al., 2018) and question answering (Bao et al., 2014; Berant et al., 2013; Das et al., 2017; Yih et al., 2015; Xu et al., 2016). Aligning language-specific knowledge bases supports NLP applications with more comprehensive commonsense reasoning (Lin et al., 2019; Li et al., 2019; Yeo et al., 2018). These applications could be particularly valuable where massive corpora for Large Language Model training are unavailable, but previous work on knowledge base construction has primarily focused on resource-rich languages (Bollacker et al., 2007; Mendes et al., 2012; Speer et al., 2016).

In this paper, we propose a novel cross-lingual mapping approach for constructing knowledge bases for low-resource languages. Our approach includes two main steps: First, we propose LeNS-Align, a novel word alignment technique that combines lexical alignment, named-entity recognition, and semantic alignment to produce high-quality word alignments over parallel text. Second, we use these word alignments to map triples from existing large-scale English knowledge bases to low-resource languages. We leverage the projected KBs to enhance multilingual language models' question answering capabilities by adapting two methods for question answering over a knowledge graph, QA-GNN (Yasunaga et al., 2021) and RGCN (Schlichtkrull et al., 2018).

Our evaluation focuses on four South African languages: isiZulu, isiXhosa, Sepedi, and SeSotho. These languages are the most spoken among the Nguni and Sotho-Tswana languages, which are the two main groups of Niger-Congo B languages in South Africa (Statistics South Africa, 2022). The agglutinative properties of these languages pose particular challenges for KB construction. While our application is tailored to these languages, LeNS-Align can be generalized to other low-resource languages, provided sufficient parallel text and basic NLP tools are available.

We created machine translated test sets from three existing QA datasets for zero-shot QA evaluation: CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), and QALD-M (Usbeck et al., 2018; Perevalov et al., 2022). Our results show that utilizing Graph Neural Networks (GNNs) to augment the mT5 language model (Xue et al., 2021) with the created knowledge bases leads to consistent improvements in QA performance across languages and datasets.

Our main contributions are: (1) LeNS-Align, a novel word alignment technique that combines lexical alignment, named entity recognition, and semantic alignment to produce high-quality word alignments from parallel text. (2) Knowledge-enhanced QA in low-resource languages: We show that the projected knowledge bases can be leveraged to enhance a multilingual language model's question answering capabilities through a GNN-based architecture. Comprehensive error analysis and ablation studies demonstrate our approach' robustness. By providing a method to construct and utilize knowledge bases in low-resource settings, we aim to facilitate more inclusive and diverse language technologies, ultimately enhancing NLP capabilities for underrepresented languages.

2 Related Work

2.1 Knowledge Base Construction

Prior to the advent of pretrained language models (PLMs), knowledge base construction relied on rule-based systems and multi-staged information extraction pipelines (Auer et al., 2007; Suchanek et al., 2007). These approaches, while effective for their time, lacked the flexibility and adaptability offered by modern PLM-based methods (Carlson et al., 2010; Lehmann et al., 2015). Specifically, PLM-based approaches enable dynamic adaptation to new domains and languages through techniques such as few-shot learning (Brown et al., 2020) and cross-lingual transfer (Schuster et al., 2019), capabilities that were not feasible with traditional rule-based systems.

Recent work has focused on multilingual Knowledge Graph (KG) embeddings for cross-KG alignment and link prediction. State-of-the-art methods like MTransE (Chen et al., 2017) and RM-GAN (Zhu et al., 2021) produce unified embedding spaces enabling link prediction in a target KG based on aligned knowledge from other KGs. While these approaches achieve strong performance on high-resource languages (with reported accuracy improvements of 15-20% over traditional methods), their effectiveness diminishes significantly for low-resource languages due to data sparsity (Chen et al., 2017, 2021; Sun et al., 2020). Our work specifically addresses this gap by introducing techniques optimized for low-resource scenarios.

2.2 Cross-lingual Knowledge Projection

Cross-lingual knowledge projection transfers knowledge from resource-rich to low-resource languages. Recent approaches include unified representation models like PRIX-LM (Zhou et al., 2022) and XLENT (El-Kishky et al., 2021), which report F1 scores of 75-80% on entity alignment tasks for major European languages. However, their performance drops by 20-30% when applied to morphologically rich languages like those in our study. Our approach achieves 85-90% accuracy on knowledge projection for morphologically complex African languages, and works with parallel corpora as small as 131k sentences, whereas previous approaches typically require 1M+ parallel sentences.

2.3 Cross-lingual Question Answering over Knowledge Graphs

Cross-lingual question answering aims to answer questions using a knowledge graph for questions in multiple languages, often different from the KG language. Typically, the QA model is trained on data and associated KG in a high-resource language, then adopted for zero-shot cross-lingual QA (Hakimov et al., 2017; Zhou et al., 2021; Zhang et al., 2023). We consider a scenario where English knowledge is projected to low-resource languages and used in a zero-shot setting to answer questions

Language Pair	Sentences	English Tokens	Target Language Tokens
English-isiZulu	232k	4.1m	2.9m
English-isiXhosa	219k	3.9m	2.7m
English-Sepedi	131k	2.8m	2.2m
English-SeSotho	164k	3.2m	2.4m

Table 1: Parallel Corpus Statistics

in these languages.

2.4 Graph Neural Networks for Question Answering

GNNs have shown effectiveness in modeling graph-based data for various NLP tasks (Yasunaga et al., 2017; Zhang et al., 2018; Yasunaga and Liang, 2020). Recent works have explored using GNNs to reason over entity graphs from supporting documents (Cao et al., 2019; Tian et al., 2021; Ma et al., 2021). Another approach uses external KB as an information source to answer questions (Feng et al., 2020). The QA-GNN (Yasunaga et al., 2021) approach jointly models language and KG components, integrating textual aspects with structured KG information. Relational Convolutional Graph Networks (RCGN) address challenges in highly multi-relational data in knowledge bases, excelling at link prediction and entity classification tasks (Schlichtkrull et al., 2018). Our work applies these GNN-based QA advances in a multilingual, low-resource context, demonstrating that knowledge-enhanced language models using GNNs can improve QA performance in languages with limited KB coverage.

3 Cross-lingual Knowledge Base Projection

3.1 Parallel Corpora

To obtain high-quality word alignments, we constructed a multilingual parallel corpus for isiZulu, isiXhosa, Sepedi, SeSotho, and English. We sourced text data primarily from South African government websites and reputable news outlets, employing a semi-automatic sentence alignment and cleaning pipeline with manual verification to ensure high alignment accuracy. The pipeline includes web-crawling, cleaning, and sentence alignment components, with manual intervention for handling misalignments and errors. Table 1 presents the statistics of the resulting parallel corpus. See Appendix A for details.

Language Pair	FastAlign	NER Align	Semantic Align
English-isiZulu	0.14	0.19	0.33
English-isiXhosa	0.19	0.24	0.35
English-Sepedi	0.13	0.20	0.32
English-SeSotho	0.17	0.23	0.34

Table 2: Alignment Error Rate scores

3.2 Word Alignment with LeNS-Align

LeNS-Align integrates three complementary information sources, each contributing to the final alignment quality:

Lexical Alignment We use FastAlign (Dyer et al., 2013) with optimized hyperparameters (iterations=10, optimization threshold=10^{-4}, p0=0.98) to establish baseline lexical correspondence between the languages by aligning words based on lexical co-occurrence. FastAlign is a log-linear reparameterization of IBM Model 2, which uses the grow-diagonal-final-and (GDFA) heuristic (Koehn et al., 2007) for symmetrizing alignments.

NER-based Alignment We use named-entity recognition (NER) models to project cross-lingual entities in our parallel datasets. We train NER models for each of the languages in our data (see Appendix B for details). Given English-Target parallel sentences, we run the English and Target NER models and align predicted named entities which match by entity type. We only align names of people, organisations, and locations.

Semantic Alignment We utilize mT5 (Xue et al., 2021) to capture deeper semantic relationships, generating contextual embeddings across languages. Word pairs are considered semantically aligned if their cosine similarity exceeds 0.8, a threshold determined through empirical validation on a development set of 1,000 manually aligned sentence pairs. To handle cases where a single word in the source language aligns with multiple words in the target language, we introduce a mechanism to keep track of the different contexts in which the alignments occur, by storing the source sentence as a context representation alongside the aligned word pairs. This allows us to disambiguate and select the most appropriate alignment based on the specific context during the knowledge projection phase (see §3.4).

```
Input: Lexical alignments $A_l$, Named entity alignments
       $A_n$, Semantic alignments $A_s$, AER values $AER_l$,
       $AER_n$, $AER_s$, Threshold $\tau$, English sentences
       $E$, Target language sentences $T$
Output: Combined alignments $A_w$
1  Calculate the alignment weights (AW):
     $AW_l = 1 - AER_l$, $AW_n = 1 - AER_n$,
     $AW_s = 1 - AER_s$
2  Initialize an empty set of combined alignments $A_w$
3  for $(w_1, w_2) \in A_l$, $e_s \in E$ and $t_s \in T$ do
4      calculate the combined probability $P$:
5      $P = AW_l$;
6      $P = P + AW_n$ if $(w_1, w_2) \in A_n$;
7      $P = P + AW_s$ if $(w_1, w_2) \in A_s$;
8      if $P \geq \tau$ then
9          $A_w \leftarrow (w_1, w_2, P, e_s, t_s)$
10     end
11 end
12 return $A_w$
```

Algorithm 1: LeNS-Align

Language	ConceptNet Triples	DBpedia Triples	Total
isiZulu	214k	464k	678k
isiXhosa	212k	461k	673k
Sepedi	226k	448k	674k
SeSotho	223k	443k	666k

Table 3: Number of KB triples mapped by our cross-lingual projection

	Correct Triples	Incorrect Triples
isiZulu ConceptNet	87.52%	12.48%
isiZulu DBpedia	90.12%	9.88%
Sepedi ConceptNet	88.56%	11.44%
Sepedi DBpedia	89.88%	10.12%

Table 4: Human evaluation results of the constructed knowledge bases.

3.3 Combining Alignments

LeNS-Align integrates alignments from lexical, named entity recognition, and semantic alignments using a novel weighted combination approach. Algorithm 1 presents the process. The algorithm assigns weights to each alignment type based on its Alignment Error Rate (AER). To estimate AER scores, we use a sample dataset of parallel sentence pairs from the English and target language corpora, and manually align the words to create reference alignments. We applied each alignment approach to the sample dataset to calculate the AER scores, shown in Table 2.

LeNS-Align processes each word pair from the lexical alignments (line 3), calculating a combined probability by summing the weights of alignments present in each method (lines 4-7). If this probability exceeds a predefined threshold τ, the word pair is added to the final alignment set A_w along with its probability and the corresponding English and target language sentences for context (lines 8-10). The threshold τ serves as a quality control mechanism, ensuring that only high-confidence alignments are included in the final set. This helps to filter out potential noise and improve the overall accuracy of the projected knowledge.

By assigning higher weights to alignments that achieved better evaluation scores, LeNS-Align prioritizes more accurate word alignments in the final set. This approach allows us to leverage the strengths of each alignment method while mitigating their individual weaknesses. The inclusion of sentence context (e_s and t_s) in the final alignment set is important for the subsequent cross-lingual knowledge projection phase. During projection, when an English word is mapped to multiple target language words, this stored context is used to disambiguate and select the most appropriate target word based on the specific context of the knowledge base triple being projected. This combined approach results in a robust set of alignments that forms the foundation for our cross-lingual knowledge projection process.

3.4 Cross-lingual Projection

With high-quality word alignments obtained through LeNS-Align, we proceed to the cross-lingual projection of knowledge base triples. This process involves mapping English (subject, predicate, object) triples to the target low-resource languages. Our projection method consists of the following steps.

Predicate Translation For each English (subject, predicate, object) triple, we first translate the predicate. We use a manually curated set of translations for the most common predicates in our knowledge bases. This ensures accurate and consistent translation of relationship types across languages.

Entity Mapping For the subject and object entities, we employ a context-aware retrieval process as outlined in Algorithm 2. The first step is to retrieve all candidate alignments for a given English entity. This is done using the `GetCandidateAlignments` function, which searches through our LeNS-Align results (A_w) and returns all target language words that have been aligned with the input English entity, along with their alignment probabilities and sentence contexts. We then compare the context

	isiZulu ConceptNet		Sepedi ConceptNet		isiZulu DBpedia		Sepedi DBpedia	
Model	**MRR**	**Hit@10**	**MRR**	**Hit@10**	**MRR**	**Hit@10**	**MRR**	**Hit@10**
TransE	0.36	0.41	0.38	0.43	0.44	0.48	0.39	0.43
ComplEx	0.48	0.494	0.483	0.495	0.48	0.56	0.52	0.57
RotatE	**0.501**	**0.53**	**0.514**	**0.542**	**0.512**	**0.58**	**0.54**	**0.594**

Table 5: Link Prediction Evaluation Results: Mean Reciprocal Rank (MRR) and Hit@10 scores for KG embedding models on the projected knowledge bases using manually verified test sets

```
   Input: English entity $e$, Knowledge Base
          triple context $c$, Alignments $A_w$
   Output: Mapped entity in target language
          $e_t$
1  $C \leftarrow$ GetCandidateAlignments($e$, $A_w$);
2  $s_{max} \leftarrow 0$;
3  for $(e_t, P, e_s, t_s) \in C$ do
4      $s \leftarrow$ ComputeContextSimilarity($c$, $e_s$);
5      if $s > s_{max}$ then
6          $s_{max} \leftarrow s$;
7          $e_{best} \leftarrow e_t$;
8      end
9  end
10 return $e_{best}$
```

Algorithm 2: Context-Aware Entity Mapping

of the knowledge base triple with the stored sentence contexts from our alignments. The target language word with the highest context similarity is selected as the mapped entity. For ConceptNet triples, we use the English Wiktionary definitions of the subject and object entities as the context. For DBpedia triples, we utilize the entity descriptions as the context. To compute context similarity (line 4), we generate embeddings for both the KB triple context and the stored alignment contexts using our fine-tuned mT5 model. We then calculate the cosine similarity between these embeddings.

Triple Construction Once we have the translated predicate and the mapped subject and object entities, we construct the projected triple in the target language:

$$(s_{en}, p_{en}, o_{en}) \rightarrow (s_t, p_t, o_t) \tag{1}$$

where s_{en}, p_{en}, o_{en} are the English subject, predicate, and object, and s_t, p_t, o_t are their corresponding translations in the target language.

Confidence Score We assign a confidence score to each projected triple based on the alignment probabilities and context similarities:

$$confidence = \frac{P_s + P_o}{2} \times sim_s \times sim_o \tag{2}$$

Where P_s and P_o are the alignment probabilities for the subject and object, and sim_s and sim_o are their respective context similarities. This process is repeated for each triple in the English knowledge base, resulting in a projected knowledge base for each target language. The confidence scores can be used to filter or rank the projected triples based on their estimated reliability.

4 Knowledge Base Evaluation

We apply our cross-lingual projection method to map subsets of ConceptNet and DBpedia to isiZulu, isiXhosa, Sepedi, and SeSotho. We focused on the top 35 relations from DBpedia and SPO triples from ConceptNet where both subject and object entities were English terms. This selection ensures a fair comparison between the two knowledge bases and focuses on the most informative relations.

Table 3 details the size of the newly constructed knowledge bases for each language. The slight variations in KB sizes across languages (e.g., 678k triples for isiZulu vs. 666k for SeSotho) can be attributed to differences in the availability of parallel text and the effectiveness of our alignment method for each language pair. These differences highlight the challenges of knowledge projection in diverse low-resource settings.

4.1 Human Evaluation

We conducted a human evaluation of the projected knowledge bases for isiZulu and Sepedi. Two native speakers of each language evaluated the accuracy of a sample of 2500 triples each from ConceptNet and DBpedia. The evaluators verified the translation of the subject, predicate, and object from the English knowledge bases. They were instructed to mark a triple as correct only if all three elements were accurately translated and the relationship remained valid in the target language.

Table 4 presents the results of this evaluation, showing over 85% of evaluated triples judged as correct across both KBs and languages. DBpedia triples showed slightly higher accuracy (90.12% for isiZulu, 89.48% for Sepedi) compared to ConceptNet triples (87.52% for isiZulu, 88.56% for Sepedi), possibly due to DBpedia's more structured relations.

4.2 Link Prediction

To further evaluate the quality and coherence of the constructed knowledge bases, we performed a link prediction task using knowledge graph embedding models. We evaluated three KG embedding models: TransE, ComplEx, and RotatE which interpret relations as translations, complex-valued embeddings to handle binary relations, and rotations, respectively. The models were trained on 90% of the triples from the KGs, tested on the remaining 10% that were randomly sampled, and evaluated on the manually verified triples.

The results in Table 5 shows strong performance across all four projected KBs, with RotatE achieving the best results. This suggests that our projection method preserves meaningful relationships in the target languages. These results demonstrate that our cross-lingual projection approach produces coherent and semantically rich knowledge graphs in the target low-resource languages.

4.3 LeNS-Align Analysis

Our analysis shows that lexical alignment contributes approximately 45% to the final alignment decisions, with particular strength in handling frequent vocabulary items. It has an accuracy of 86.5% for words appearing more that 100 times in the corpus. NER-based alignment is particularly crucial for handling proper nouns and technical terms, contributing 30% to final alignment decisions, with 92.3% accuracy for named entities.

5 Zero-shot Question Answering

5.1 KBQA

We use mT5 (Xue et al., 2021) as our base language model, with continued pre-training on Nguni and Sotho-Tswana language corpora to improve coverage of target languages (See Appendix C). For CommonsenseQA and OpenBookQA, we use ConceptNet as knowledge base, while for QALD-M, we use DBpedia. This choice aligns with the nature of the questions in each dataset: CommonsenseQA and OpenBookQA focus on general knowledge, while QALD-M contains more factual questions that align well with DBpedia's structure.

We implement two methods for question answering over a knowledge graph using GNNs: (1) QA-GNN (Yasunaga et al., 2021) and RGCN (Schlichtkrull et al., 2018). Both methods were adapted for our multilingual setting and projected knowledge bases.

First we obtain knowledge graph embeddings by training GNNs on the projected knowledge graph in each language. For this, we implemented Multi-Hop Graph Relation Network (MHGRN) proposed by Feng et al. (2020), and used mT5 as the text encoder. Following Yasunaga et al. (2021), to implement QA-GNN we first use mT5 to encode the question-answer (QA) context, and retrieve a subgraph from KB using the approach from Feng et al. (2020). We design a joint graph using the QA context and the retrieved subgraph, where the QA context is connected to the topic entities within the KBs subgraphs. The attention-based GNN module performs reasoning on the joint graph.

We implement relational graph convolutional network (RGCN) (Schlichtkrull et al., 2018) as a graph autoencoder, and adapt it for question answering. RGCN is trained on the projected knowledge graphs to obtain embeddings. We use mT5 to encode the QA context, retrieve a subgraph, and join the QA context to the subgraph as a node. We then fed the joint subgraph into RGCN and get updated embeddings for all nodes in the subgraph. We calculate the similarity scores between the QA context node and other entities in the subgraph, and retrieve the top ranked entity as answer to the question.

5.2 Experimental Setup

For each target language (isiZulu, isiXhosa, Sepedi, and SeSotho), we sampled and machine-translated 3k question-answer pairs from three datasets: CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), and QALD-M (Usbeck et al., 2018; Perevalov et al., 2022). We restructured the data into a fill-in-the-blank format for zero-shot evaluation, transforming questions like "A yard is made up of what?" to "A yard is made up of __". This format allows for a more direct evaluation of the model's ability to leverage the projected knowledge bases.

We compare our knowledge-enhanced models (QA-GNN and RGCN) with two baselines: (1)

Dataset	Method	isiZulu		isiXhosa		Sepedi		SeSotho	
		Acc.	Hit@5	Acc.	Hit@5	Acc.	Hit@5	Acc.	Hit@5
CommonsenseQA	mT5	0.61	0.61	0.59	0.60	0.56	0.59	0.55	0.58
	mT5+KG	0.65	0.68	0.63	0.65	0.68	0.71	0.65	0.68
	RGCN	0.69	0.73	0.67	0.70	0.69	0.77	0.65	0.72
	QA-GNN	**0.72**	**0.77**	**0.70**	**0.73**	**0.75**	**0.80**	**0.71**	**0.76**
OpenBookQA	mT5	0.59	0.60	0.57	0.58	0.53	0.57	0.52	0.56
	mT5+KG	0.63	0.66	0.61	0.63	0.66	0.70	0.64	0.66
	RGCN	0.68	0.74	0.66	0.69	0.69	0.75	0.66	0.72
	QA-GNN	**0.75**	**0.78**	**0.72**	**0.74**	**0.79**	**0.80**	**0.75**	**0.78**
QALD-M	mT5	0.60	0.63	0.58	0.61	0.54	0.59	0.54	0.57
	mT5+KG	0.69	0.71	0.65	0.67	0.66	0.69	0.64	0.67
	RGCN	0.73	0.74	0.70	0.72	0.71	**0.75**	0.67	**0.72**
	QA-GNN	**0.80**	**0.81**	**0.77**	**0.78**	**0.73**	**0.75**	**0.69**	**0.72**

Table 6: Combined Main Results: Test Accuracy and Hit@5 for different methods across datasets and languages.

Dataset	Method	isiZulu			isiXhosa			Sepedi			SeSotho		
		Full	Ablated		Full	Ablated		Full	Ablated		Full	Ablated	
			NER	Sem		NER	Sem		NER	Sem		NER	Sem
CommonsenseQA	mT5+KG	0.65	0.55	0.57	0.63	0.52	0.54	0.68	0.57	0.59	0.65	0.62	0.63
	RGCN	0.69	0.62	0.65	0.67	0.61	0.64	0.69	0.64	0.66	0.65	0.63	0.64
	QA-GNN	**0.72**	0.69	0.70	**0.70**	0.67	0.68	**0.75**	0.72	0.73	**0.71**	0.71	0.70
OpenBookQA	mT5+KG	0.63	0.56	0.58	0.61	0.54	0.56	0.66	0.58	0.61	0.64	0.64	0.63
	RGCN	0.68	0.63	0.66	0.66	0.61	0.64	0.69	0.65	0.67	0.66	0.66	0.65
	QA-GNN	**0.75**	0.71	0.73	**0.72**	0.69	0.70	**0.79**	0.72	0.76	**0.75**	0.71	0.73
QALD-M	mT5+KG	0.69	0.63	0.66	0.65	0.59	0.62	0.66	0.57	0.63	0.64	0.56	0.62
	RGCN	0.73	0.66	0.68	0.70	0.64	0.67	0.71	0.67	0.69	0.67	0.65	0.66
	QA-GNN	**0.80**	0.75	0.76	**0.77**	0.73	0.75	**0.73**	0.70	0.72	**0.69**	0.66	0.67

Table 7: Combined Main Results and Ablation Study: Test Accuracy for different methods across datasets and languages. 'Full' represents results with complete LeNS-Align, 'NER' shows results with NER component removed, 'Sem' shows results with Semantic Alignment component removed. %Drop shows the maximum performance drop from Full to either ablation.

Vanilla mT5, with no KB access, serving as a baseline to enable assessing the impact of knowledge injection; (2) mT5+KG: mT5 augmented by fine-tuning on verbalized triples from the projected KBs. This baseline helps isolate the impact of the graph structure in our GNN-based approaches. All experiments are conducted in a zero-shot setting to specifically investigate the impact of injected knowledge on QA performance.

5.3 Results and Analysis

Table 6 reports test accuracy and Hit@5 results for CommonsenseQA, OpenBookQA, and QALD-M across the four languages. For isiZulu and Sepedi, we report results from manually translated datasets. The results show consistent improvements in accuracy and Hit@5 scores across all datasets and languages when mT5 is augmented with the projected knowledge bases. This demonstrates the effectiveness of our knowledge projection approach in enhancing zero-shot QA capabilities for low-resource languages. The mT5+KG baseline shows improvements over vanilla mT5, indicating that even simple knowledge injection techniques can be beneficial.

QA-GNN outperforms RGCN in all QA tasks across the four languages. This can be attributed to QA-GNN's ability to jointly reason over both the question text and the KG structure, allowing it to better leverage the contextual information in the questions. We observe better relative performance for Sepedi and SeSotho on CommonsenseQA and OpenBookQA, while isiZulu and isiXhosa show higher proficiency on QALD-M. This may be attributed to differences in knowledge base sizes and the nature of the questions in each dataset. Performance on QALD-M is generally higher than on CommonsenseQA and OpenBookQA. This could be due to the more factual nature of QALD-M questions, which may align better with the structured knowledge in DBpedia.

6 Ablation Study

To understand the relative importance of LeNS-Align's components, we conducted ablation experiments by removing two key components: the Named Entity Recognition (NER) system and the semantic alignment mechanism. These experiments reveal how each component contributes to the overall system performance.

Table 7 presents the results of these ablations across all target languages and evaluation datasets. The removal of the NER component, shown in Table 7, leads to performance decreases across all experimental conditions. The mT5+KG model shows the highest sensitivity to NER removal, with accuracy dropping by more than 15 percentage points for isiZulu on CommonsenseQA and more than 17 percentage points for isiXhosa. The impact is particularly pronounced on QALD-M tasks, where accuracy decreases by 8 to 13 percentage points across all languages.

The RGCN model demonstrated moderate resilience to NER removal, with performance decreases ranging from 2 to 7 percentage points. The QA-GNN architecture proved most robust, maintaining relatively stable performance even without NER. For instance, on CommonsenseQA, QA-GNN's accuracy dropped by only 3 percentage points for isiZulu and isiXhosa, while showing minimal degradation for Sepedi and SeSotho.

Table 7 reveals a different pattern when removing the semantic alignment component. The performance impact is generally less severe than NER removal, with accuracy decreases ranging from 2 to 4 percentage points across models. The mT5+KG model again shows the highest sensitivity, particularly for Nguni languages, where accuracy dropps by 8 percentage points for isiZulu and 9 points for isiXhosa on CommonsenseQA.

The differential impact between NER and semantic alignment removal suggests their distinct roles in the system. NER appears crucial for maintaining overall system performance, particularly for tasks requiring precise entity identification and handling. The larger performance drops observed with NER removal, especially in entity-centric QALD-M tasks, highlight its fundamental importance to the pipeline.

In contrast, the more modest impact of removing semantic alignment indicates that while this component contributes to system performance, other components can partially compensate for its absence. The graph-based architectures (RGCN and QA-GNN) show particular resilience to both types of ablation, suggesting their ability to leverage graph structure helps maintain performance even with reduced input quality.

These findings reveal the complementary yet distinct roles of LeNS-Align's components. While both NER and semantic alignment contribute to system performance, NER plays a more critical role in enabling accurate cross-lingual knowledge projection and question answering.

7 Conclusion

This paper introduced LeNS-Align, a novel approach for constructing knowledge bases in low-resource languages through the integration of lexical alignment, named entity recognition, and semantic alignment techniques. Empirical results across four low-resource South African languages validate LeNS-Align's effectiveness across multiple evaluation dimensions. Human evaluation demonstrated over 85% accuracy in projected triples, while link prediction results indicated strong semantic coherence in the projected knowledge bases. In downstream question answering tasks, our approach improved accuracy by up to 17% over baseline methods across all target languages and datasets. Ablations revealed the important role of named entity recognition in producing high-quality cross-lingual alignments, with its removal particularly affecting entity-centric tasks. Nguni languages showed higher sensitivity to component removal compared to Sotho-Tswana languages, suggesting the need for language-family-specific adaptations. Our results demonstrate that combining multiple alignment strategies with neural architectures can significantly improve cross-lingual knowledge projection and question answering capabilities for underserved languages.

Limitations

The effectiveness of LeNS-Align is constrained by its dependence on parallel corpora and the performance of NER models, limiting applicability in low-resource languages where these resources are not available. Error propagation can lead to cascading inaccuracies from the alignment and knowledge projection stages to downstream tasks with knowledge augmentation. The morphological complexity of agglutinative languages also poses challenges, as the current approach may not fully account for divergence across languages. Furthermore, manual verification steps during the development process help to ensure quality but hinder scalability across languages and larger knowledge bases. Finally, the projected knowledge base were developed in specific cultural context that might not align with that of the speakers of the target low-resource languages.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850). Sello Ralethe is supported by the Hasso Plattner Institute for Digital Engineering, through the HPI Research School at the University of Cape Town.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. 2014. Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 967–976. The Association for Computer Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Kurt D. Bollacker, Robert P. Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 1962–1963. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2306–2317. Association for Computational Linguistics.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press.

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth. 2021. Cross-lingual entity alignment with incidental supervision. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 645–658. Association for Computational Linguistics.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1511–1517. ijcai.org.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 358–365. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza*

Hotel, Atlanta, Georgia, USA, pages 644–648. The Association for Computational Linguistics.

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten south african languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3698–3703. European Language Resources Association (ELRA).

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10424–10430. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1295–1309. Association for Computational Linguistics.

Sherzod Hakimov, Soufian Jebbara, and Philipp Cimiano. 2017. AMUSE: multilingual semantic parsing for question answering over linked data. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, volume 10587 of *Lecture Notes in Computer Science*, pages 329–346. Springer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Shiyang Li, Jianshu Chen, and Dian Yu. 2019. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach. *CoRR*, abs/1909.09743.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics.

Jiangtao Ma, Duanyang Li, Yonggang Chen, Yaqiong Qiao, Haodong Zhu, and Xuncai Zhang. 2021. A knowledge graph entity disambiguation method based on entity-relationship embedding and graph structure embedding. *Comput. Intell. Neurosci.*, 2021:2878189:1–2878189:11.

Xin Mao, Wenting Wang, Huimin Xu, Yuanbin Wu, and Man Lan. 2020. Relational reflection entity alignment. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1095–1104. ACM.

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1813–1817. European Language Resources Association (ELRA).

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2018. Never-ending learning. *Commun. ACM*, 61(5):103–115.

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *J. Web Semant.*, 51:1–19.

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022. Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In *16th IEEE International Conference on Semantic Computing, ICSC 2022, Laguna Hills, CA, USA, January 26-28, 2022*, pages 229–234. IEEE.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max

Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3795–3805. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Statistics South Africa. 2022. Census 2022. `https://census.statssa.gov.za/assets/documents/2022/P03014_Census_2022_Statistical_Release.pdf`.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 222–229. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Luogeng Tian, Bailong Yang, Xinli Yin, Kai Kang, and Jing Wu. 2021. Multipath cross graph convolution for knowledge representation learning. *Comput. Intell. Neurosci.*, 2021:2547905:1–2547905:13.

Ricardo Usbeck, Ria Hari Gusmita, Axel-Cyrille Ngonga Ngomo, and Muhammad Saleem. 2018. 9th challenge on question answering over linked data (QALD-9) (invited paper). In *Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th - 9th, 2018*, volume 2241 of *CEUR Workshop Proceedings*, pages 58–64. CEUR-WS.org.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Michihiro Yasunaga and Percy Liang. 2020. Graph-based, self-supervised program repair from diagnostic feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10799–10808. PMLR.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir R. Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 452–462. Association for Computational Linguistics.

Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang. 2018. Machine-translated knowledge transfer for commonsense causal reasoning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2021–2028. AAAI Press.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1321–1331. The Association for Computer Linguistics.

Chen Zhang, Yuxuan Lai, Yansong Feng, Xingyu Shen, Haowei Du, and Dongyan Zhao. 2023. Cross-lingual question answering over knowledge base as reading comprehension. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2394–2407. Association for Computational Linguistics.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5429–5435. ijcai.org.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2205–2215. Association for Computational Linguistics.

Wenxuan Zhou, Fangyu Liu, Ivan Vulic, Nigel Collier, and Muhao Chen. 2022. Prix-lm: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5412–5424. Association for Computational Linguistics.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5822–5834. Association for Computational Linguistics.

Renbo Zhu, Meng Ma, and Ping Wang. 2021. RAGA: relation-aware graph attention networks for global entity alignment. In *Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part I*, volume 12712 of *Lecture Notes in Computer Science*, pages 501–513. Springer.

A Parallel Corpus Creation

Existing parallel datasets for the four low-resource languages are mostly automatically constructed by scrapping webpages and then using a language identification model to align sentences across the languages. This approaches is fiddled with errors and also depends on the accuracy of the language identification model.

In order to obtain high-quality word alignments, we constructed a multilingual parallel corpus for isiZulu, isiXhosa, Sepedi, SeSotho and English using text data sourced primarily from South African government websites[1] and news websites. We employed a semi-automatic sentence alignment and cleaning pipeline with manual verification to ensure high alignment accuracy. The pipeline includes a web-crawling component to scrape text data from identified websites, and a cleaning and alignment component with manual intervention for handling cases of sentence misalignment and undetected errors.

Sentence alignment was performed on a per-web page basis, with manual intervention utilized to correct errors that arose from inconsistent sentence counts. We implemented a verification process by randomly selecting sentence pairs and comparing them across languages, ensuring that they were semantically equivalent. In cases where errors were identified, we implemented manual corrections through minor edits or adding missing sentences in one or more languages.

Table 1 shows the statistics of the parallel corpus for the four different language pairs, giving the number of sentences and number of words for each language pair.

B Named Entity Recognition Models

For the Named Entity Recognition (NER) component of LeNS-Align, we developed specialized NER models for each target language (isiZulu, isiXhosa, Sepedi, and SeSotho) as well as English. These models play a crucial role in identifying and aligning named entities across languages, enhancing our knowledge projection process.

B.1 Model Architecture

We implemented a Bidirectional Long Short-Term Memory (Bi-LSTM) architecture for our NER models. This choice was motivated by the Bi-LSTM's ability to capture contextual information from both

[1] https://www.gov.za/

directions in a sequence, which is particularly useful for NER tasks.

The model architecture consists of:

1. An embedding layer to convert input tokens into dense vector representations
2. A Bi-LSTM layer to process the embedded sequences
3. A time-distributed dense layer with softmax activation for entity classification

B.2 Training Data

We used the NCHLT Text Resource Development dataset (Eiselen and Puttkammer, 2014) for training our NER models. This dataset provides annotated text for several South African languages, including our target languages. The dataset includes annotations for person names, organization names, and location names.

B.3 Training Process

The models were trained using the following hyperparameters:

- Embedding dimension: 100
- LSTM hidden units: 100
- Batch size: 32
- Number of epochs: 10
- Optimizer: Adam with learning rate 0.001

We used an 80-10-10 split for training, validation, and test sets.

B.4 Model Performance

The performance of our NER models on the test set for each language is summarized in Table 8.

Language	**Precision**	**Recall**	**F1-Score**
English	0.92	0.90	0.91
isiZulu	0.83	0.81	0.82
isiXhosa	0.82	0.80	0.81
Sepedi	0.81	0.79	0.80
SeSotho	0.79	0.77	0.78

Table 8: NER Model Performance

These results demonstrate the performance of our NER models across all target languages, with English showing the highest performance. The lower performance for the African languages can be attributed to the more complex morphological structures and less training data compared to English.

C mT5 Continued Pre-training

To enhance the performance of our question answering system for low-resource South African languages, we performed continued pre-training of the multilingual T5 (mT5) model (Xue et al., 2021). This process involved further training the pre-trained model on our target languages to better capture their linguistic nuances and structures.

C.1 Model Selection and Hardware

We chose the mT5-large model as our starting point due to its strong performance on multilingual tasks and its capacity to capture complex linguistic patterns. The mT5-large model has approximately 1.2 billion parameters, offering a good balance between model capacity and computational feasibility for continued pre-training.

For our computational infrastructure, we utilized a Google Cloud Compute Engine instance with the following specifications:

- Machine Type: a2-ultragpu-2g
- GPUs: 2 x NVIDIA A100 80GB
- Memory: 340GB

C.2 Pre-training Data

We compiled a diverse corpus for each target language:

- News articles from major South African news websites
- Government documents and reports
- Wikipedia articles (where available)
- Multilingual Educational materials
- Multilingual short stories

The size of the pre-training corpora varied by language, as shown in Table 9.

C.3 Pre-processing

We applied the following pre-processing steps:

- Text cleaning (removing HTML tags, standardizing punctuation)

Language	Tokens (millions)	Unique Words
isiZulu	87	1.9M
isiXhosa	83	1.5M
Sepedi	45	0.3M
SeSotho	51	0.6M

Table 9: Size of Continued Pre-training Corpora

- Tokenization using the SentencePiece model from the original mT5
- Removal of sentences with more than 50% non-alphabetic characters
- Deduplication at the document level

C.4 Continued Pre-training Process

We continued pre-training using the original mT5 objective: a denoising task where the model must reconstruct randomly masked spans of input text. The training was performed on a Google Cloud Compute Engine instance with two NVIDIA A100 80GB GPUs.

- Batch size: 16 per GPU (32 total, with gradient accumulation steps of 8, resulting in an effective batch size of 256)
- Learning rate: 5e-5 with linear decay and 2000 warmup steps
- Number of epochs: 3 (approximately 75,000 steps)
- Maximum sequence length: 512 tokens
- Masked span length: Average of 3 tokens, determined by a Poisson distribution ($\lambda = 3$)
- Masking probability: 15% of all tokens
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 8$
- Weight decay: 0.01
- Gradient clipping: 1.0

We implemented several techniques to maximize the use of our dual-GPU setup:

- Mixed precision training (FP16) to reduce memory usage and speed up computations
- Data parallelism across the two GPUs to effectively double our processing capacity
- Gradient accumulation to fine-tune our effective batch size

C.5 Integration with LeNS-Align

The continued pre-trained mT5 model was integrated into the LeNS-Align pipeline for two main purposes:

1. Generating contextual word embeddings for semantic alignment
2. Providing a strong baseline for the question answering task, which was further enhanced by the knowledge graph integration

This continued pre-trained model played an important role in bridging the gap between the original pre-trained multilingual model and the specific requirements of our low-resource language task.