Design Challenges in Low-resource Cross-lingual Entity Linking

Xingyu Fu[1]*, **Weijia Shi**[2]*, **Xiaodong Yu**[1], **Zian Zhao**[3], **Dan Roth**[1]
[1]University of Pennsylvania
[2]University of California Los Angeles
[3]Columbia University
{xingyuf2, xdyu, danroth}@seas.upenn.edu
swj0419@g.ucla.edu, zz2562@columbia.edu

Abstract

Cross-lingual Entity Linking (XEL), the problem of grounding mentions of entities in a foreign language text into an English knowledge base such as Wikipedia, has seen a lot of research in recent years, with a range of promising techniques. However, current techniques do not rise to the challenges introduced by text in low-resource languages (LRL) and, surprisingly, fail to generalize to text not taken from Wikipedia, on which they are usually trained. This paper provides a thorough analysis of low-resource XEL techniques, focusing on the key step of identifying candidate English Wikipedia titles that correspond to a given foreign language mention. Our analysis indicates that current methods are limited by their reliance on Wikipedia's interlanguage links and thus suffer when the foreign language's Wikipedia is small. We conclude that the LRL setting requires the use of outside-Wikipedia cross-lingual resources and present a simple yet effective zero-shot XEL system, QuEL, that utilizes search engines query logs. With experiments on 25 languages, QuEL shows an average increase of **25%** in gold candidate recall and of **13%** in end-to-end linking accuracy over state-of-the-art baselines.[1]

1 Introduction

Cross-lingual Entity Linking (XEL) aims at grounding mentions written in a foreign (source) language (SL) into entries in a (target) language Knowledge Base (KB), which we consider here as the English Wikipedia following Pan et al. (2017); Upadhyay et al. (2018a); Zhou et al. (2020). In Figure 1, for instance, an Odia (an Indo-Aryan language in India) mention ("Chilika Lake") is linked to the corresponding English Wikipedia entry. The XEL

---

* Both authors contributed equally to this work.
[1]Code is available at: http://cogcomp.org/page/publication_view/911. The LORELEI data will be available through LDC.
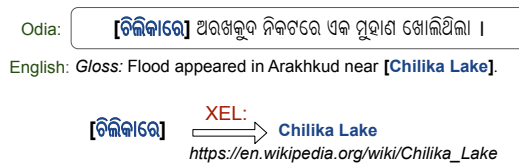

Figure 1: The EXL task: in the given sentence we link "Chilika Lake" to its corresponding English Wikipedia

task typically involves two main steps: (1) candidate generation, retrieving a list of candidate KB entries for the mention, and (2) candidate ranking, selecting the most likely entry from the candidates.

While XEL techniques have been studied heavily in recent years, many challenges remain in the LRL setting. Specifically, existing candidate generation methods perform well on Wikipedia-based dataset but fail to generalize beyond Wikipedia, to news and social media text. Error analysis on existing LRL XEL systems shows that the key obstacle is candidate generation. For example, 79.3%-89.1% of XEL errors in Odia can be attributed to the limitations of candidate generation.

In this paper, we present a thorough analysis of the limitations of several leading candidate generation methods. Although these methods adopt different techniques, we find that all of them heavily rely on Wikipedia interlanguage links[2] as their cross-lingual resources. However, small SL Wikipedia size limits their performance in the LRL setting. As shown in Figure 2, while the core challenge of LRL XEL is to link LRL entities (A) to candidates in the English Wikipedia (C), interlanguage links only map a small subset of the LRL entities that appear in both LRL Wikipedia (B) and English Wikipedia. Therefore, methods that only leverage interlanguage links ($B \cap C$) as the main source of supervision cannot cover a wide range of entities.

---

[2]https://en.wikipedia.org/wiki/Help:Interlanguage_links

For example, the Amharic Wikipedia has 14,854 entries, but only 8,176 of them have interlanguage links to English.

Furthermore, as we show, existing candidate generation methods perform well on Wikipedia-based datasets but fail to generalize to outside-Wikipedia text such as news or social media.
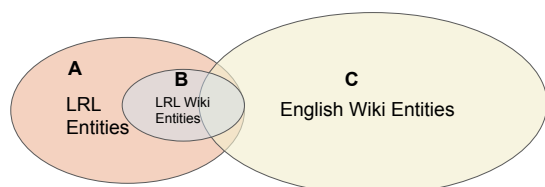

Figure 2: An illustration of Cross-lingual resources serving current XEL systems.

Our observations lead to the conclusion that the LRL setting necessitates the use of outside-Wikipedia cross-lingual resources. Specifically, we propose various ways to utilize the abundant query log from online search engines to compensate for the lack of supervision. Similar to Wikipedia, a free online encyclopedia created by Internet users, Query logs (QL) provide a free resource, collaboratively generated by a large number of users, and mildly curated. However, it is orders of magnitude larger than Wikipedia.[3] In particular, it includes all of Wikipedia cross-lingual resources as a subset since a search of an SL mention leads to the English Wikipedia entity if the corresponding Wikipedia entries are interlanguage-linked.

In this paper, the main part, Sec. 3 presents a thorough *method-wise* evaluation and analysis of leading candidate generation methods, and quantifies their limitations as a function of SL Wikipedia size and the size of the interlanguage cross-lingual resources. Based on the limitations, we analyze QuEL_CG, an improved candidate generation method utilizing QL in Sec. 4, showing that it exceeds the Wikipedia resource limit on LRL. To exhibit a *system-wise* XEL comparison, in Sec. 5, we suggest a simple yet efficient zero-shot XEL framework QuEL, that incorporates QuEL_CG. QuEL achieves an average of **25%** increase in gold candidate recall, and **13%** increase in end-to-end linking accuracy on outside-wikipedia text.

2 Related Work

2.1 Background: Wikipedia resources

Wikipedia title mappings: The mapping comes from Wikipedia interlanguage links[4] between the SL and English, and uses Wikipedia articles titles as mappings. It directly links an SL entity to an English Wikipedia entry without ambiguity.

Wikipedia anchor text mappings: A clickable text mention in Wikipedia articles is annotated with anchor text linking it to a Wikipedia entry. The following retrieving order: SL anchor text $\rightarrow SL$ Wikipedia entry $\rightarrow$ English Wikipedia entry (where the last step is done via the Wikipedia interlanguage links), allows one to build a bilingual title mapping from a SL mentions to Wikipedia English entries, resulting in a probabilistic mapping with scores calculated using total counts (Tsai and Roth, 2016).

2.2 XEL systems for Low-resource languages

We briefly survey key approaches to XEL below.

Direct Mapping Based Systems, including xlwikifier (Tsai and Roth, 2016) and xelms (Upadhyay et al., 2018a), focus on building a SL to English mapping to generate candidates. For candidate ranking, both xlwikifier and xelms combine supervision from multiple languages to learn a ranking model.

Word Translation Based Systems including Pan et al. (2017) and ELISA (Zhang et al., 2018), extract SL – English name translation pairs and apply an unsupervised collective inference approach to link the translated mention.

Transliteration Based Systems include Tsai and Roth (2018) and translit (Upadhyay et al., 2018b). translit uses a sequence-to-sequence model and bootstrapping to deal with limited data. It is useful when the English and SL word pairs have similar pronunciation.

Pivoting Based Systems including (Rijhwani et al., 2019; Zhou et al., 2019) and PBEL_PLUS (Zhou et al., 2020), remove the reliance on SL resources and use a *pivot* language for candidate generation. Specifically, they train the XEL model on a selected high-resource language, and apply it to SL mentions through language conversion.

2.3 Query Logs

Query logs have long been used in many tasks such as across-domain generalization (Rüd et al., 2011) for NER, and ontological knowledge acquisition(Alfonseca et al., 2010; Pantel et al., 2012). In the English entity linking task, Shen et al. (2015) pointed out that google query logs can be an efficient way to identify candidates. Dredze et al. (2010); Monahan et al. (2011) use search result as one of their methods for candidate generation on high resource language entity linking task. While earlier work indicates that query logs provide abundant information that can be used in NLP tasks, as far as we know, it has never been used in cross-lingual tasks as we do here. And, only search information has been used, while we suggest using Maps information too.

3 Candidate Generation Analysis

In this section, we analyze four leading candidate generation methods: `p(e|m)`, `xlwikifier`, `name_trans`, `pivoting`, and `translit`(see Table 1 for the systems used by each method) and discuss their limitations.

Method	p(e|m)	name_trans	pivoting	translit
System Usage				
xlwikifier	✓	-	-	-
xelms	✓	-	-	-
ELISA	-	✓	-	-
PBEL_PLUS	✓	-	✓	-
Cross-lingual Resource				
Title Map	✓	✓	✓	✓
Anchor Text	✓	-	✓	-
Other Lang	-	-	✓	-

Table 1: Wikipedia cross-lingual resources and systems (introduced in Section 5) used by LRL XEL candidate generation methods.

3.1 Candidate Generation Methods

Each method listed in Table 1, is discussed below along with the level of resources it requires.

p(e|m) (Tsai and Roth, 2016) creates a direct probabilistic mapping table using Wikipedia title mappings and the anchor text mappings, between SL and English. E.g., if an Oromo mention "Itoophiyaatti" is the anchor text linked to an Oromo Wikipedia entity "Itoophiyaa"[5], and "Itoophiyaa" has an interlanguage link to English Wikipedia entity "Ethiopia", then "Ethiopia" will be added as a candidate for the mention "Itoophiyaatti". Thus, `p(e|m)` follows a linking

---

[5] https://om.wikipedia.org/wiki/Itoophiyaa

---

flow: SL mention $\rightarrow$ SL Wikipedia entity $\rightarrow$ English Wikipedia entity.23

name_trans (Name Translation) as introduced in (Pan et al., 2017; Zhang et al., 2018) performs word alignment on Wikipedia title mappings, to induce a fixed word-to-word translation table between SL and English. For instance, to link the Suomi name "Pekingin tekninen instituutti" (Beijing Institute of Technology), it translates each word in the mention: ("Pekingin" – Beijing, "tekninen" – Technology, "instituutti" – Institute). At test time, after mapping each word in the given mention to English, it links the translation to English Wikipedia using an unsupervised collective inference approach.

translit (Upadhyay et al., 2018b) trains a seq2seq model on Wikipedia title mappings, to generate the English entity directly.

pivoting to a related high-resource language (HRL) (Zhou et al., 2020) attempts to generalize to unseen LRL mentions through grapheme or phoneme similarity between SL and a related HRL. Specifically, it first finds a related HRL for the SL, and learns a XEL model on the related HRL, using Wikipedia title mappings and anchor text mappings. Then it applies the model on SL mentions. If SL and related HRL share a script, showing grapheme similarity, it treats SL mentions as related HRL mentions at test time. Otherwise, if SL and related HRL have phoneme similarity it converts both SL and related HRL text into international phonetic alphabet (IPA) symbols.

3.2 Current Methods' Limitations

This section discusses four major limitations that existing methods suffer from, and quantifies these with experimental results. The results use the LORELEI dataset (Strassel and Tracey, 2016), a realistic text corpora that consist of news and social media text, all from outside-Wikipedia (see Section 5). Some tables also include a comparison with the proposed QL based candidate generation method `QuEL_CG` that we describe in Section 4. We use the definition of *gold candidate recall* in Zhou et al. (2020), which is the proportion of SL mentions that have the gold English entity in the candidate list, as the evaluation metric.

3.2.1 Shortage of Interlanguage Links

As illustrated in Figure 3 (specific numbers are in Appendix A.1) with statistics from the 2019-10-20 wikidump[6] and in Table 2 for five randomly picked
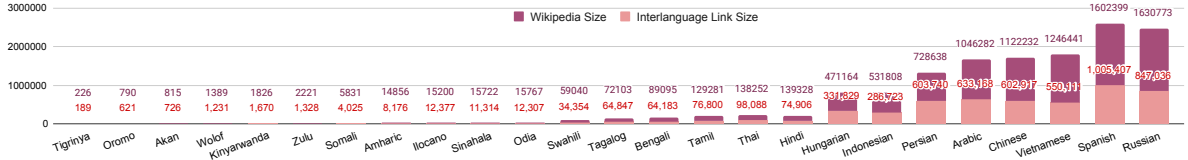
---

[6] https://dumps.wikimedia.org/

3000000

2000000

1000000

■ Wikipedia Size   ■ Interlanguage Link Size

1602399 1630773

1046282 1122232 1246441

728638

471164 531808

226 790 815 1389 1826 2221 5831 14856 15200 15722 15767 59040 72103 89095 129281 138252 139328 331829 286723 608710 633183 602917 550411 1005497 847036
189 621 726 1231 1670 1328 4025 8176 12377 11314 12307 34354 64847 64183 76800 98088 74906

0

Tigrinya Oromo Akan Wolof Kinyarwanda Zulu Somali Amharic Ilocano Sinhala Odia Swahili Tagalog Bengali Tamil Thai Hindi Hungarian Indonesian Persian Arabic Chinese Vietnamese Spanish Russian

Figure 3: SL Wikipedia size (number of articles) and Interlanguage links size (between SL and English). We define low-resource languages (LRL) as Tigrinya to Odia, and high-resource ones (HRL) as Swahili to Russian.

Lang.	Oromo	Akan	Wolof	Zulu	Somali
SL-only Entity (%)	34.68	70.93	37.54	49.77	12.93

Table 2: Proportion of gold mentions in the LORELEI corpus with English Wikipedia pages but no SL ones.

Model	Hindi	Bengali	Odia	Sinahala
translit-Wiki	24.6	23.4	13.4	8.6

Table 3: *Gold candidate recall* of translit.

low-resource languages, many LRL Wikipedias only have a few interlanguage links to the English Wikipedia. Consequently, only a few Wikipedia title mappings and anchor text mappings are accessible by all four methods, as shown in Table 1.

For p(e|m), the workflow is: SL mention → SL Wikipedia entity → English Wikipedia entity, and it breaks if one link is missing. For example, "Nawala" has an English Wikipedia page, but does not have a corresponding Sinhala page. Given its Sinhala mention[7], the interlink is missing and p(e|m) returns 0 probability for "Nawala".

For name_trans, its translation ability is limited by the tokens contained in the Wikipedia title mappings. For a SL mention, when none of its tokens ever appeared in the SL Wikipedia titles, it will not have any English translation, and thus generate no candidates. As for translit and pivoting, they will have fewer data pairs to train on and the model performance would suffer.

3.2.2 Small LRL Mention Coverage

In the LRL setting, few Wikipedia articles lead to fewer Wikipedia anchor text mappings, thus reducing the ability of current methods to cover many SL mentions. For instance, the LRL Oromo Wikipedia article for "Laayibeeriyaa"[8] has much fewer hyperlinks than the English Wikipedia article for "Liberia"[9], even though they are linked through an interlanguage link. Figures 3 and 4 show that *gold candidate recall* generally goes up with the increase in the Wikipedia size.

To evaluate Wikipedia coverage, we propose a global metric called *mention token coverage* that can be computed without the gold data. *Mention token coverage* is the percentage of SL mentions that have at least one token appearing in Wikipedia title mappings or in anchor text mappings. For example, when we consider Somali language, the mention "Shabeelada hoose" has a token "hoose" covered in Wikipedia titles, so it is counted in the *mention token coverage*. But"Soomaalieed" is not covered in Wikipedia titles or anchor text mappings, so it is not counted for *mention token coverage*. High *mention token coverage* values tend to guarantee better supervision when trained on Wikipedia. Indeed, in Figure 4, we can clearly see that the *mention token coverage* for LRLs is much smaller than that of high-resource languages (consult Figure 3 for the distinction between LRLs and HRLs).

We also compare the *mention token coverage* with *gold candidate recall* for each method in Figure 4. When a method has a *gold candidate recall* higher than *mention token coverage*, it means that the method is able to generalize beyond Wikipedia. In contrast, when *gold candidate recall* is lower than *mention token coverage*, it implies that the method is limited by Wikipedia resources.

To compare relation between *mention token coverage* and *gold candidate recall* more clearly, in Figure 5, we show the ratio of *gold candidate recall* over *mention token coverage*. Existing methods' ratio ranges between 0.31 to 1.27, with the average of 0.72. This suggests that existing methods are bounded by Wikipedia resources in most cases. This figure also shows the generalization ability of our QL-based candidate generation method QuEL_CG (introduced in Sec.4) to outside-Wikipedia mentions with an average ratio of 1.13 ranging between 0.74 and 1.92.

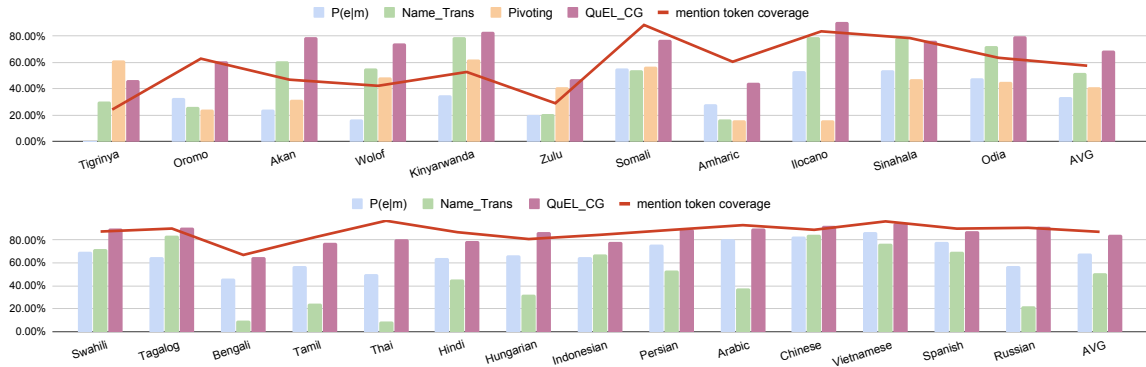---

[7] https://foursquare.com/v/nawala--%E0%B6%B1%E0%B7%80%E0%B6%BD--%E0%AE%A8%E0%AE%B5%E0%AE%B2/4eaa947bf7905c39414c250d/photos

[8] https://om.wikipedia.org/wiki/Laayibeeriyaa

[9] https://en.wikipedia.org/wiki/Liberia

Figure 4: *Gold candidate recall* (bars) on low-resource languages (top) and high-resource languages (bottom) correlates with *mention token coverage* (line) for existing approaches (excluding `QuEL_CG`). *Mention token coverage* is the percentage of SL mentions that have at least one token overlapping with a SL Wikipedia title or anchor text (that correspond to it via interlanguage link). Languages appear in increasing order of Wikipedia size.
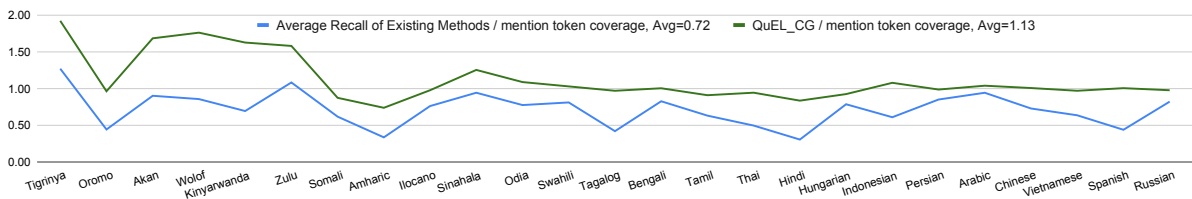

Figure 5: **Average *gold candidate recall* remains 0.5-0.8 times that of *mention token coverage* for existing approaches, but the ratio is 1.0-2.0 for `QuEL_CG`.** The lines are the ratio of *gold candidate recall* per candidate generation method divided by *mention token coverage*. Average *gold candidate recall* of existing methods: `p(e|m)`, `name_trans`, `pivoting`, is mostly limited by *mention token coverage* and cannot exceed it, with 0.72 average ratio. However, our proposed `QuEL_CG` can reach recall up to 2 times that of *mention token coverage*, with average ratio of 1.13 on all languages, tested to be statistically significant with p-value $< 0.01\%$.

3.2.3 `translit` Data Requirements

`translit` suffers from the inability to satisfy several of its data requirements. Typically, transliteration models need many training data pairs, which is hard to get using LRL Wikipedia title pairs. Also, they require SL mentions and gold English entities to have word-by-word mappings. Table 3 shows results of `translit` trained on name pairs from Wikipedia title mappings (`translit-Wiki`). We only provide results for languages for which the model has been released; this sample already shows clearly that the performance of `translit` drops significantly on LRLs.

3.2.4 `pivoting` Prerequisites

While `pivoting` does not suffer from insufficient cross-lingual resources between SL and English, it is limited by resources between related HRL and English. More importantly, it is limited the availability of related HRL that is similar enough to SL. `pivoting` learns through grapheme or phoneme similarity. However, grapheme similarity is not enough since not every LRL has a related

HRL that uses the same scripts. In these cases `pivoting` uses phoneme similarity and maps strings to international phonetic alphabet (IPA) symbols. For example, (Zhou et al., 2020) uses Epitran (Mortensen et al., 2018) to convert strings to IPA symbols, but Epitran only supports 55 of the 309 Wikipedia languages, and covers only 8 out of 12 low-resource languages in the LORELEI corpus.

As in Figure 4, `pivoting` gains from the language conversion compared with `p(e|m)`, but the increase varies among languages, depending on the choice of related HRL. Most importantly, the related HRL cannot replace SL and language conversion may limit on the linking ability. For example, the language pair Oromo-Indonesian(grapheme) has same scripts but `pivoting` still suffers from low-resource on Oromo.

4 Improved Candidate Generation: `QuEL_CG`

As mentioned in Section 3, Wikipedia's cross-lingual resources are not enough for XEL to per-

form well in low-resource settings. We argue that outside-Wikipedia resources are essential to compensate for the lack of supervision. We suggest that search engines query logs provide an excellent resource in this situation, since it is a (very large) super-set of Wikipedia's cross-lingual data but covers even more; as pointed out in the Introduction, it is a collaboratively generated, mildly curated, resource and its effective size is a function of the number of SL native speakers using the search engine. Thus, query logs can help the candidate generation process map SL mentions to English Wikipedia candidates even when they are not covered by Wikipedia interlanguage links.

We propose an improved candidate generation method, `QuEL_CG`, that uses query log mapping files. We obtain a high-quality candidate list through directly searching SL mentions in query logs and a query-based pivoting method. While we can choose any search engines, we use here Google search[10] and Google maps[11]. `QuEL_CG` also runs in conjunction with `p(e|m)` to cover the cases where query log mapping is not robust.

4.1 Query Logs as Search Results

As the first step, we search the morphologically normalized SL mention in the Google search engine (implementation details in Appendix A.2) and retrieve a list of web-page results. We pick top k (k is usually 1 or 5) Wikipedia web-page results P_k. Note that if the searched result is SL Wikipedia article and it has an interlanguage link to English, we convert it to the corresponding English one through the link, and then mark the corresponding English entity as a candidate. When the SL mention is a geopolitical or location entity, we search its normalized mention using Google Map. Since it only returns English surface of the location instead of Wikipedia articles, we further search the resulted English surface in Google search using the same procedure described above.

4.2 Query-based Pivoting

We also conduct language-indifferent pivoting using query logs. Note that the pivoting methods described below are different from `pivoting` in Section 3.

Some LRLs have high-resource languages they are similar to (e.g. Sinhala to Hindi and Tigrinya to Amharic). In order to exploit the similarities between an LRL and HRL without having to choose one good HRL, we use query logs for pivoting. We first follow the same search steps described above to get top k Wikipedia web-pages results P_k. Finally, we continue the same process on the new pivoted mentions. A special case here is language-specific pivoting on selected language pairs. We use a simple utf-8 converter to translate SL mention into a related, but higher-resource language, such as Odia to Hindi, and then run the candidate generation process described above on the pivoted mention.

5 Experiments: System Comparison

Given the analysis of the key candidate generation process in Section 3, this section moves to study its implications on the overall performance of different LRL XEL systems. We first propose our LRL XEL framework, `QuEL`, by combining `QuEL_CG` with a zero-shot candidate ranking module. Our experimental goal is to compare all systems on both outside-Wikipedia data and Wikipedia data. We further analyze the entity distribution and entity type on the linking results. In addition, an ablation study is demonstrated in Appendix A.3.

5.1 Datasets

Dataset details are reported in Appendix A.1. **LORELEI dataset** (Strassel and Tracey, 2016) is a realistic and challenging dataset that includes news and social media such as twitter. We divide its 25 languages into LRL and HRL as in Figure 3. Entities in LORELEI are of four types: geopolitical entities (GPE), locations (LOC), persons (PER) and organizations (ORG). The dataset provides a specific English KB that mentions are linked to; we processed the original dataset to link to the English Wikipedia instead. Our processed gold labels will be available along with LORELEI dataset[12]. Given a KB entity, we link it to Wikipedia if the KB provides a Wikipedia link. For a PER or ORG entity without Wikipedia link, we use its KB-provided English information, e.g. name and description, to search for Wikipedia entry, and manually check the correctness. Otherwise, we do not include this entity and remove any mentions linking to it in the EDL dataset. We process these types differently because PER and ORG entities only compose around

---

5% of the gold entities.

Wikipedia-based dataset collected by (Tsai and Roth, 2016) is built upon Wikipedia anchor text mappings. All languages in this dataset are high-resource ones as defined in Figure 3.

5.2 System Comparison

We compare the **supervised** SOTA systems, `xlwikifier`, `xelms`, `ELISA`, `PBEL_PLUS`, that use candidate generation methods analyzed earlier, with a new, QL-based system, that we present below. Implementation details are in Appendix A.4

5.3 A QL-based XEL: `QuEL`

Given the limitations discussed in Sec. 3 we propose a new XEL system, `QuEL`, that uses `QuEL_CG` (Sec. 4) along with the following zero-shot candidate ranking module. Given a candidate list C_m (the output of `QuEL_CG` on SL mention m), `QuEL` uses the multilingual BERT (Devlin et al., 2018) to score the candidates against m. Specifically, for each candidate $c \in C_m$ it computes a score $W(c, m)$ that measures "relatedness" between m and c. It then picks the candidate with the highest score as its output. In case of a tie, we break it by following the candidate selection order, from Google search results, to Google Map results, to `p(e|m)` candidates. We explain below the components of $W(c, m)$.

Candidate Multiplicity Weight A candidate can be suggested by multiple sources–Google search, Google Map search, query-based pivoting, or `p(e|m)`. `QuEL` prefers candidates generated by multiple sources. We define candidate c multiplicity weight as: $W_{Source}(c) = Num_{Source}(c)$, the number of sources that generate c.

Contextual Disambiguation. `QuEL` uses Multilingual BERT (M-BERT) (Devlin et al., 2018) for multilingual embeddings, to compute the similarity of the context of the mention m and the candidate's context in the English Wikipedia. We denote by $W_{context}(c, m)$ the cosine similarity between m's context embedding and c's context embedding (see details in Appendix A.5). Finally, the score for candidate c is $W(c, m) = W_{source}(c) \cdot W_{context}(c, m)$, and we select the most likely entity: $e = \arg\max_{c \in C_m} W(c, m)$ as output.

5.4 Entity Linking Results

Comprehensive evaluations of both the LORELEI and the Wikipedia based datasets are shown in Figures 6 and 7 and in Table 4. (Scores that correspond

to the figures are reported in Appendix A.6.) Note that gold candidate recall on `xlwikifier` and `xelms` are identical because they use the same candidate generation module, `p(e|m)`.

`QuEL` is shown to significantly improve over existing approaches on both datasets, especially on the more difficult LORELEI dataset where it improves on almost all the languages and shows an average of **25%** increase in gold candidate recall and of **13%** in linking accuracy. On the Wikipedia-based dataset, `QuEL` shows an average increase of **4%** in gold candidate recall, while reaching the SOTA on linking accuracy. Importantly, most other systems, (`ELISA`, `xlwikifier` and `xelms`) use *supervised* ranking modules.

System	Gold Candidate Recall	Linking Accuracy
	LORELEI Dataset	
`xlwikifier`	52.54	46.58
`xelms`	52.54	48.65
`ELISA`	50.52	43.91
`PBEL_PLUS`	38.36	30.38
`QuEL`	**78.21**	**61.40**
	Wikipedia-based Dataset	
`xlwikifier`	79.40	63.51
`xelms`	79.40	**68.73**
`ELISA`	47.82	41.22
`QuEL`	**83.54**	66.16

Table 4: *Gold candidate recall* and *linking accuracy* comparison averaged over all languages.

6 Analysis

Table 4 shows huge performance gaps between the two datasets using the SOTA baseline `xelms`. The 20% percentage difference in both metrics proves that the LORELEI dataset is more difficult due to having more outside-Wikipedia mentions, and more focus on low-resource languages. One exception is `ELISA`, which has lower performance on Wikipedia-based dataset. We believe it fails to cover many Wikipedia mentions because it does not use Wikipedia anchor text mappings.

Similarly, when we consider the same language performance on the two datasets (see Figures 6 and 7), e.g., Tamil and Thai, the LORELEI dataset appears harder to deal with. However, `QuEL` achieves similar results on both datasets, and also brings the gold candidate recall for LRL much closer to that of HRL. Another important observation is that `QuEL` performs significantly better on the LORELEI dataset. It suggests that our proposed `QuEL_CG` addresses the outside-Wikipedia coverage problem well by exploiting the query logs. To understand why `QuEL` exceeds baselines largely on
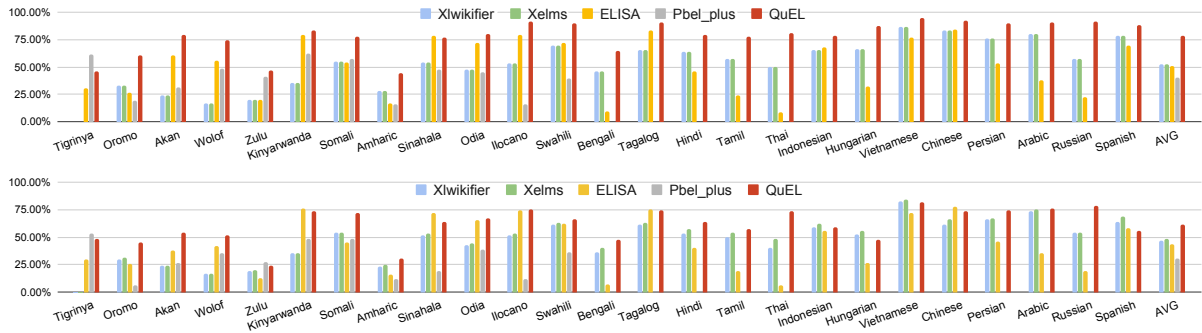
Figure 6: End-to-end XEL gold candidate recall(top) and linking accuracy(bottom) on the LORELEI dataset sorted by Wikipedia size in ascending order. Specific scores are reported in Tables 11 and 12 in Appendix A.6.
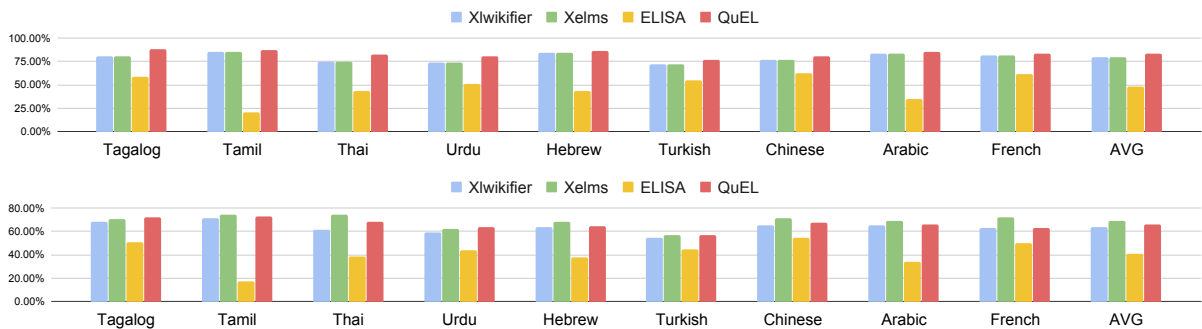

Figure 7: End-to-end XEL gold candidate recall(top) and linking accuracy(bottom) on the Wikipedia-based dataset sorted by Wikipedia size in ascending order. Specific scores are reported in Table 13 in Appendix A.6.

System	Akan	Oromo	Zulu	Wolof	Somali
	Gold Candidate Recall				
xelms	23.9	33.2	19.8	16.9	55.1
ELISA	60.7	26.2	20.4	55.5	54.3
PBEL_PLUS	31.7	19.1	40.8	48.5	57
QuEL	**79.0**	**60.6**	**47.1**	**74.4**	**77.4**
	Linking Accuracy				
xelms	23.9	31.4	19.8	16.6	54.5
ELISA	38.00	25.6	12.4	42.2	45.0
PBEL_PLUS	26.5	5.9	27.7	35.5	48.6
QuEL	**53.8**	**45.4**	**23.8**	**51.8**	**72.1**

Table 5: *Gold candidate recall* and *linking accuracy* for SL-only entities in correspondence to Table 2.

Lang.	System	Accuracy (%)			
		GPE	LOC	PER	ORG
Oromo	xelms	33.0	3.6	6.2	6.3
	ELISA	37.0	11.6	12.5	**11.9**
	PBEL_PLUS	53.0	40.5	6.4	0.0
	QuEL	**57.5**	30.4	17.2	4.2
Zulu	xelms	25.6	0.0	37.5	5.7
	ELISA	11.9	2.5	31.2	5.7
	PBEL_PLUS	24.6	6.2	38.1	8.7
	QuEL	**27.8**	**10.8**	**50.0**	**12.6**
Somali	xelms	55.0	28.6	21.4	42.1
	PBEL_PLUS	44.7	14.3	**25.6**	**75.0**
	ELISA	50.1	0.0	16.7	15.0
	QuEL	**71.7**	**57.1**	33.3	65.0

Table 6: *Linking accuracy* on different types of entities.

LORELEI dataset, we analyze it on entity resource and type distribution as below.

Entity Resource. Considering the insufficient Wikipedia interlanguage links for LRL in Table 2, we investigate whether QuEL_CG helps in this situation. In Table 5, QuEL shows 6.3% to 27.4% improvement in gold candidate recall, indicating that it can effectively perform XEL without the Wikipedia cross-lingual resource, leading to a significant improvement for LRL XEL.

Entity Type. Table 6 shows the evaluation on all four types of entities. We observe that QuEL improves more on GPE and LOC entities, than on

PER and ORG entities. We believe the improvement is brought by Google Map query logs.

7 Conclusion

This work provides a thorough analysis of existing LRL XEL techniques, focusing on the step of generating English candidates for foreign language mentions. The analysis identifies the inherent lack of sufficient inter-lingual supervision signals as a key shortcoming of the current approach. This leads to proposing a rather simple method that leverages query logs, that are highly effective in addressing

these challenges. Given that our experiments show a 25% increase in the candidate generation, one future research direction is to improve candidate ranking in LRL by incorporating coherence statistics and entity types. Moreover, given the effectiveness of query logs, we believe it can be applied to other cross-lingual tasks like relation extraction and Knowledge Base completion.

Acknowledgments

References

Enrique Alfonseca, Marius Pasca, and Enrique Robledo-Arnuncio. 2010. Acquisition of instance attributes via labeled and related instances. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 58–65.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285.

Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale, and Arnold Jung. 2011. Cross-lingual cross-document coreference with entity linking. In *TAC*.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1946–1958.

Patrick Pantel, Thomas Lin, and Michael Gamon. 2012. Mining entity types from query logs via user intent modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 563–571. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: Using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 965–975. Association for Computational Linguistics.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3273–3280.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual Wikification Using Multilingual Embeddings. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chen-Tse Tsai and Dan Roth. 2018. Learning Better Name Translation for Cross-Lingual Wikification. In *Proc. of the Conference on Artificial Intelligence (AAAI)*.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018a. Joint Multilingual Supervision for Cross-Lingual Entity Linking. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shyam Upadhyay, Jordan Kodner, and Dan Roth. 2018b. Bootstrapping Transliteration with Constrained Discovery for Low-Resource Languages. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Boliang Zhang, Ying Lin, Xiaoman Pan, Di Lu, Jonathan May, Kevin Knight, and Heng Ji. 2018. Elisa-edl: A cross-lingual entity extraction, linking and localization system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 41–45.

Shuyan Zhou, Shruti Rijhawani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *arXiv preprint arXiv:2003.01343*.

Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. Towards zero-resource cross-lingual entity linking. *arXiv preprint arXiv:1909.13180*.

A Appendices

A.1 Dataset Statistics

We show the dataset statistics calculated from 2019-10-20 Wikidump[13] as below, demonstrating the Wikipedia size (article number), interlanguage link size (between SL and English), and test data mention number for every language, on both LORELEI dataset and Wikipedia-based dataset.

Language	Interlang./Wiki	# Test Mentions
Tagalog	64,847/72,103	1,075
Tamil	76,800/129,281	1,075
Thai	98,088/138,252	11,380
Urdu	128,227/154,103	1,390
Hebrew	193,391/267,243	16,137
Turkish	244,882/354,767	13,795
Chinese	612,335/1,122,232	11,252
Arabic	633,168/1,046,282	1,0647
French	1,398,118/2,221,709	2,637

Table 8: Overview for Wikipedia-based dataset. Wiki/Interlang. refers to corresponding Wikipedia and English interlanguage link size.

Language	Interlang./Wiki	#Test Mentions
Tigrinya	189/226	3174
Oromo	621/790	2576
Akan	726/815	462
Wolof	1,231/1389	302
Zulu	1,328/2221	1071
Kinyarwanda	1,670/1826	521
Somali	4,025/5,831	884
Amharic	8,176/14,856	1157
Sinahala	11,314/15,722	673
Odia	12,307/15,767	2079
Ilocano	12,377/15,200	1274
Swahili	34,354/59,040	1251
Bengali	64,183/89,095	1266
Tagalog	64,847/72,103	1050
Hindi	74,906/139,328	814
Tamil	76,800/129,281	1157
Thai	98,088/138,252	1122
Indonesian	286,723/531,808	1376
Hungarian	331,829/471,164	1059
Vietnamese	550,111/1,246,441	990
Chinese	612,335/1,122,232	1157
Persian	603,740/728,638	877
Arabic	633,168/1,046,282	1188
Russian	847,036/1,630,773	1205
Spanish	1,005,407/1,602,399	711

Table 7: Overview for LORELEI dataset. Wiki/Interlang. refers to corresponding Wikipedia and English interlanguage link size.

---

[13]https://dumps.wikimedia.org

A.2 Implementation Details of `QuEL_CG`

To perform our improved candidate generation method `QuEL_CG`, we first conduct morphological normalization on the SL mention before querying. Then, we use the normalization output as search input. We use Google search [14] and Google Map [15]. We can also customize the search input for better results on extremely low-resource languages (Odia and Illocano) as below.

Morphological normalization. Language-specific Morphological normalization is a basic process for all candidate generation methods. An entity may have different surface forms in the document, which makes candidate generation difficult. To cope with this issue, several operations including removing, adding, or replacing suffixes and prefixes are conducted as a prior process.

Customize search input. To better retrieve Wikipedia pages as search results and ignore other web-page results, "wiki" or "[Country of SL]" can be appended to the original search input.

A.3 Ablation Study

We now quantify the effects of each component in our candidate generation method and show the results in Table 9.

Google Map. Our model is default to use the Google Map cross lingual resource. We test the effect of adding supervision from this QL.

Google top1. In everywhere that takes the Google query log (QL) results, take only the first Wikipedia page result that is in source or target language as candidate.

Google top5. Similar to Google top1, take the top 5 Wikipedia page results as candidates. We can see that Google top1 and top5's effects are language dependent.

`p(e|m)`. We test whether adding the `p(e|m)` module would help in linking performance. To better show the results, `p(e|m)` is added under the setting of using QL and Google Map KB, without adding other modules.

Pivoting. Pivoting here refers to our query-based pivoting, different from `pivoting` in Section 3. We picked two low-resource languages: Odia and Tigrinya, to explore the pivoting effect and show results in Table 10. On Odia, language-specific pivoting skill is used. Since we know in prior that Odia

and Hindi are similar while the latter has much more resource, a simple utf8-converter is used to transform Odia to Hindi, and then runs the Hindi mention through our whole system. On Tigrinya, a language-indifferent pivoting skill is used. After getting QL results, besides using Google top1 or top5, we further pick Wikipedia page results that are in any other language, such as Amharic or Scots that have similar scripts, but with richer cross-lingual supervision then Tigrinya.

We further examined the effect of transliteration models using trained models (Upadhyay et al., 2018b) specifically on Sinhala and Odia, with bilingually mapped Wikipedia titles as supervision. We also used Google transliteration resource for Odia mentions. However, no increase on linking accuracy is observed, and the absolute increase in gold candidate recall is less than 0.5%. Since we only studied on Sinhala and Odia, maybe the transliteration resource is useful on other languages.

Tables 9 and 10 show that we added a lot of value beyond the use of Google search – simply using google search without adding other parts of our candidate generation methods does not have good linking results. Indeed, incorporating online search engine query logs effectively to XEL is highly non-trivial. In this context, it is important to note that all existing methods make heavy use of Wikipedia, and therefore using QL as a cross-lingual resource is as fair. Moreover, as our results show, the use of Wikipedia allows existing systems to perform well only on Wikipedia data, which is uninteresting for all practical purposes. As shown in Figure 6, Tables 11 and 12, our method works well outside Wikipedia!

	Odia (%)		Tigrinya (%)	
	Accuracy	Recall	Accuracy	Recall
`QuEL` w/o pivot.	66	78.6	45.3	**46.4**
`QuEL`	**66.7**	**79.3**	**45.7**	**46.4**

Table 10: Ablation study on 2 low-resource languages to examine effect of pivoting techniques. To better show the difference, a simple setting, Google top1 + `p(e|m)`, is used for each language.

A.4 Implementation Details of Compared Systems

For `xlwikifier`[16], we use different versions of candidate ranking on the two datasets. Since the

---

[14] https://developers.google.com/custom-search/v1/overview

[15] https://cloud.google.com/maps-platform

[16] https://github.com/cttsai/illinois-cross-lingual-wikifier

	Akan (%)		Thai (%)		Tigrinya (%)	
	Accuracy	**Cand. Recall**	**Accuracy**	**Cand. Recall**	**Accuracy**	**Cand. Recall**
Google top1 w/o Google Map	53.4	57.9	73.5	74.6	31.7	31.9
Google top1	54.7	61	**74**	77.1	**44.9**	46.4
Google top5	54	79	73.8	79.5	37.2	48.7
Google top1 + p(e|m)	**55.1**	61	73.8	78.8	45.3	46.4
Google top5 + p(e|m)	53.8	**79**	73.5	**80.9**	36.3	**48.7**

	Oromo (%)		Somali (%)		Oria (%)	
	Accuracy	**Cand. Recall**	**Accuracy**	**Cand. Recall**	**Accuracy**	**Cand. Recall**
Google top1 w/o Google Map	40	43.8	67.8	71.4	47.6	55.3
Google top1	43.9	50.1	71.8	76.3	59.2	70
Google top5	41.8	55.5	70.7	80.6	56.5	76.6
Google top1 + p(e|m)	**45.4**	57.2	**72.1**	77.4	**66**	78.6
Google top5 + p(e|m)	42.7	**60.6**	71.2	**81.5**	64.6	**83.5**

Table 9: Ablation study on 6 low-resource languages that examines each candidate generation component for end-to-end linking accuracy (left) and gold candidate recall(right). Candidate number is below 5 in most languages and varies between 2-9. Our method as default includes Google Map module.

Wikipedia-based dataset provides training data, we use its provided version of candidate ranking. However, the LORELEI dataset has no training data, and thus no candidate ranking model can be trained. We just pick the first candidate as the result. For a comparison purpose, since xelms uses the same candidate generation module and provides better candidate ranking, xlwikifier is close to and mostly up-bounded by xelms results.

For **xelms**[17], we use trained ranking modules on most languages when available, and use the zero-shot version of ranking module for the rest of languages (Akan and Kinyarwanda).

For **ELISA**, we access the system using the API [18] directly provided by its authors, and call the GET/entity_linking/{identifier} function.

For **PBEL_PLUS**[19], we test this approach only on low-resource languages on the LORELEI dataset, because it generates candidates through pivoting to a related high-resource language, and it does not make sense to pivot a already high-resource language to other languages.

A.5 Implementation Details of Our System

During candidate ranking, for a mention m in a document D, we get the sentence s_m where m appears and computes its contextualized embedding $v_m = \text{M-BERT}(e, s_m)$. For each $c \in C_m$, we retrieve a list of sentences $S_c = \{s_1, s_2, ..., s_n\}$ that contains the candidate entity c in its corresponding Wikipedia page's summary. The contextualized embedding for c is denoted by v_c:

$$v_c = \frac{1}{|S_c|} \sum_{i=1}^{n} \text{M-BERT}(c, s_i)$$

Note that we picked two representative languages: Odia and Ilocano, for which we have additional LORELEI provided monolingual text, and trained the M-BERT model (Devlin et al., 2018) using their Wikipedia data along with LORELEI text. We did not use pre-trained M-BERT [20] on all languages because many low-resource languages are not supported, and for the supported ones the performance increase is much less than that of models trained with LORELEI text plus Wikipedia data. This experiment serves to show the gain one could get from additional supervision and, at the same time, highlights the results we show when M-BERT is not available, which is more realistic.

A.6 Comprehensive Evaluation

This section includes comprehensive evaluation on XEL systems.

Language	Method	Accu	Rec@5	Rec@n
Tamil	xlwikifier	49.8	57.4	57.4
	xelms	53.8	57.4	57.4
	ELISA	19.6	24.1	24.4
	QuEL	**58.2**	**73.6**	**76.6**
Zulu	xlwikifier	19.6	19.8	19.8
	xelms	19.8	19.8	19.8
	ELISA	12.4	17.9	20.4
	PBEL_PLUS	27.7	33.7	40.8
	QuEL	**23.8**	**41.2**	**47.1**
Akan	xlwikifier	23.9	23.9	23.9
	xelms	23.9	23.9	23.9
	ELISA	38	60.5	60.7
	PBEL_PLUS	26.5	28.0	31.7
	QuEL	**53.8**	**78.1**	**79**
Amharic	xlwikifier	23.3	28.2	28.2
	xelms	24.6	28.2	28.2
	ELISA	16.4	16.7	16.8
	PBEL_PLUS	11.7	11.9	16.0
	QuEL	**30.7**	**43.8**	**44.7**
Hindi	xlwikifier	53.5	63.9	63.9
	xelms	57.4	63.9	63.9
	ELISA	40.3	43.4	45.8
	QuEL	**63.6**	**74.4**	**79**
Indonesian	xlwikifier	59.2	65.3	65.3
	xelms	**62.2**	65.3	65.3
	ELISA	56	64	67.7
	QuEL	60	**73.2**	**74.6**
Spanish	xlwikifier	63.9	78.1	78.1
	xelms	**68.4**	78.1	78.1
	ELISA	57.8	68.3	69.8
	QuEL	56	**81.5**	**87.9**
Arabic	xlwikifier	73.3	80.4	80.4
	xelms	75.1	80.4	80.4
	ELISA	35.5	37.3	37.9
	QuEL	**75.6**	**84**	**90.2**
Swahili	xlwikifier	61.3	69.6	69.9
	xelms	63.4	69.6	69.6
	ELISA	62	71.4	72.2
	PBEL_PLUS	36.2	37.3	39.3
	QuEL	**66.3**	**76.2**	**76.2**
Wolof	xlwikifier	16.6	16.9	16.9
	xelms	16.6	16.9	16.9
	ELISA	42.2	52.2	55.5
	PBEL_PLUS	35.5	42.2	48.5
	QuEL	**51.8**	**66.1**	**66.1**
Vietnamese	xlwikifier	**82.4**	86.9	86.9
	xelms	84.1	86.9	86.9
	ELISA	72.1	76.7	76.9
	QuEL	81.3	**91.3**	**95**
Thai	xlwikifier	40	50.1	50.1
	xelms	48.3	50.1	50.1
	ELISA	6.2	9.1	9.1
	QuEL	**73.8**	**79.4**	**79.5**
Bengali	xlwikifier	36.5	46.4	46.4
	xelms	40.7	46.4	46.4
	ELISA	7.3	9.4	9.9
	QuEL	**47.4**	**61.6**	**65**
Tagalog	xlwikifier	61.4	65.3	65.3
	xelms	63.2	65.3	65.3
	ELISA	**75.3**	82.3	83.6
	QuEL	74.1	**88.5**	**90.4**
Hungarian	xlwikifier	52.5	66.4	66.4
	xelms	**55.8**	66.4	66.4
	ELISA	26.3	31.6	32.2
	QuEL	47.7	**78.1**	**87.2**

Table 11: Quantitative evaluation results on 25 languages on LORELEI dataset. Accu is linking accuracy, Rec@n is gold candidate recall, with n ranging between 2 to 9 for QuEL and 100 for PBEL_PLUS. Rec@5 is gold candidate recall if we reserve only top5 candidates by the ranking score.

Language	Method	Accu	Rec@5	Rec@n
Chinese	xlwikifier	61.4	83.2	83.2
	xelms	66.4	83.2	83.2
	ELISA	**77.3**	83.6	84.5
	QuEL	73.8	**89.8**	**92.4**
Persian	xlwikifier	66.1	76.1	76.1
	xelms	67	76.1	76.1
	ELISA	46.1	53	53.4
	QuEL	**74.7**	**84.6**	**89.5**
Russian	xlwikifier	53.9	57.4	57.4
	xelms	54.1	57.4	57.4
	ELISA	19.1	20.8	22.2
	QuEL	**78.6**	**87.7**	**91.2**
Oromo	xlwikifier	29.7	33.2	33.2
	xelms	31.4	33.2	33.2
	ELISA	25.6	26.1	26.2
	PBEL_PLUS	5.9	20.6	24.2
	QuEL	**45.4**	**57.2**	**57.2**
Tigrinya	xlwikifier	0	0	0
	xelms	0	0	0
	ELISA	30	30.4	37
	PBEL_PLUS	53.4	56.7	61.6
	QuEL	45.7	46.4	46.4
Sinhala	xlwikifier	51.9	54.1	54.1
	xelms	52.9	54.1	54.1
	ELISA	72	**77.7**	**78.2**
	PBEL_PLUS	19.2	26.7	47.3
	QuEL	64.1	72.8	76.8
Kinyarwanda	xlwikifier	35.1	35.1	35.1
	xelms	35.1	35.1	35.1
	ELISA	**75.9**	79.2	79.2
	PBEL_PLUS	48.5	51.4	62.0
	QuEL	73.6	**83.4**	**83.4**
Ilocano	xlwikifier	52.0	53.2	53.2
	xelms	53.2	53.2	53.2
	ELISA	74.2	77.4	79.5
	PBEL_PLUS	12.3	13.3	16.1
	QuEL	**74.9**	**84.9**	**91.1**
Odia	xlwikifier	42.6	47.6	47.6
	xelms	44.29	47.6	47.6
	ELISA	65.1	71.8	72.3
	PBEL_PLUS	39.1	42.0	45.5
	QuEL	**66.7**	**79.2**	**79.7**
Somali	xlwikifier	54.5	55.1	55.1
	xelms	54.5	55.1	55.1
	ELISA	45	53.1	54.3
	PBEL_PLUS	48.6	54.5	57.0
	QuEL	**71.2**	**80.7**	**81.5**

Table 12: Quantitative evaluation results on 25 languages on LORELEI dataset (continued). Accu is linking accuracy, Rec@n is gold candidate recall, with n ranging between 2 to 9 for QuEL and 100 for PBEL_PLUS. Rec@5 is gold candidate recall if we reserve only top5 candidates by the ranking score.

Language	Method	Accu	Rec@5	Rec@n
Arabic_wiki	xlwikifier	65.2	83.4	83.4
	xelms	**69.2**	83.4	83.4
	ELISA	34.1	34.5	34.9
	QuEL	66.1	**85.2**	**85.5**
French_wiki	xlwikifier	62.7	81.1	81.9
	xelms	**71.8**	81.6	81.9
	ELISA	50.3	59	61.2
	QuEL	63.2	**82.5**	**83.5**
Hebrew_wiki	xlwikifier	63.5	84.4	84.9
	xelms	**68.4**	84.9	84.39
	ELISA	37.39	42.4	43.2
	QuEL	64.6	**86.3**	**86.8**
Tamil_wiki	xlwikifier	71.5	85.6	85.8
	xelms	**74.1**	85.8	85.8
	ELISA	16.9	20	20.5
	QuEL	72.8	**87.4**	**87.5**
Thai_wiki	xlwikifier	73.36	75.1	75.3
	xelms	**74.5**	75.3	75.3
	ELISA	38.9	42.6	43.5
	QuEL	68.1	**82.4**	**82.4**
Tagalog_wiki	xlwikifier	68	80.3	80.4
	xelms	70.5	80.4	80.4
	ELISA	50.5	57.5	58.8
	QuEL	**72.3**	**88.6**	**88.7**
Turkish_wiki	xlwikifier	54.5	72.1	72.5
	xelms	56.8	72.4	72.5
	ELISA	44.5	52.2	54.7
	QuEL	**57.1**	**76.1**	**76.5**
Urdu_wiki	xlwikifier	59.5	73.2	73.5
	xelms	62.4	73.5	73.5
	ELISA	43.6	50.1	51
	QuEL	**63.8**	**80.4**	**80.7**
Chinese_wiki	xlwikifier	64.9	76.8	76.9
	xelms	**71.2**	76.9	76.9
	ELISA	54.3	60.4	62.6
	QuEL	67.4	**80.3**	**80.3**

Table 13: Quantitative evaluation results on 9 languages on Wikipedia-based dataset. Accu is linking accuracy, Rec@n is gold candidate recall, with n ranging between 2 to 9 for QuEL and 100 for PBEL_PLUS. Rec@5 is gold candidate recall if we reserve only top5 candidates by the ranking score.