

Entity Linking in 100 Languages

Jan A. Botha
Google Research
jabot@google.com

Zifei Shan
Google Research
zifeishan@gmail.com

Daniel Gillick
Google Research
dgillick@google.com

Abstract

We propose a new formulation for *multilingual entity linking*, where language-specific mentions resolve to a language-agnostic Knowledge Base. We train a dual encoder in this new setting, building on prior work with improved feature representation, negative mining, and an auxiliary entity-pairing task, to obtain a single entity retrieval model that covers 100+ languages and 20 million entities. The model outperforms state-of-the-art results from a far more limited cross-lingual linking task. Rare entities and low-resource languages pose challenges at this large-scale, so we advocate for an increased focus on zero- and few-shot evaluation. To this end, we provide **Mewsl-9**, a large new multilingual dataset¹ matched to our setting, and show how frequency-based analysis provided key insights for our model and training enhancements.

1 Introduction

Entity linking (EL) fulfils a key role in grounded language understanding: Given an ungrounded entity mention in text, the task is to identify the entity’s corresponding entry in a Knowledge Base (KB). In particular, EL provides grounding for applications like Question Answering (Férvy et al., 2020b) (also via Semantic Parsing (Shaw et al., 2019)) and Text Generation (Puduppully et al., 2019); it is also an essential component in knowledge base population (Shen et al., 2014). Entities have played a growing role in representation learning. For example, entity mention masking led to greatly improved fact retention in large language models (Guu et al., 2020; Roberts et al., 2020).

But to date, the primary formulation of EL outside of the standard monolingual setting has been *cross-lingual*: link mentions expressed in one language to a KB expressed in another (McNamee et al., 2011; Tsai and Roth, 2016; Sil et al., 2018).

The accompanying motivation is that KBs may only ever exist in some well-resourced languages, but that text in many different languages need to be linked. Recent work in this direction features progress on low-resource languages (Zhou et al., 2020), zero-shot transfer (Sil and Florian, 2016; Rijhwani et al., 2019; Zhou et al., 2019) and scaling to many languages (Pan et al., 2017), but commonly assumes a single primary KB language and a limited KB, typically English Wikipedia.

We contend that this popular formulation limits the scope of EL in ways that are artificial and inequitable.

First, it artificially simplifies the task by restricting the set of viable entities and reducing the variety of mention ambiguities. Limiting the focus to entities that have English Wikipedia pages understates the real-world diversity of entities. Even within the Wikipedia ecosystem, many entities only have pages in languages other than English. These are often associated with locales that are already underrepresented on the global stage. By ignoring these entities and their mentions, most current modeling and evaluation work tend to side-step underappreciated challenges faced in practical industrial applications, which often involve KBs much larger than English Wikipedia, with a much more significant zero- or few-shot inference problem.

Second, it entrenches an English bias in EL research that is out of step with the encouraging shift toward *inherently multilingual* approaches in natural language processing, enabled by advances in representation learning (Johnson et al., 2017; Pires et al., 2019; Conneau et al., 2020).

Third, much recent EL work has focused on models that rerank entity candidates retrieved by an alias table (Férvy et al., 2020a), an approach that works well for English entities with many linked mentions, but less so for the long tail of entities and languages.

¹<http://goo.gle/mewsl-9-dataset>

To overcome these shortcomings, this work makes the following key contributions:

- Reformulate entity linking as inherently multilingual: link mentions in 104 languages to entities in WikiData, a language-agnostic KB.
- Advance prior dual encoder retrieval work with improved mention and entity encoder architecture and improved negative mining targeting.
- Establish new state-of-the-art performance relative to prior cross-lingual linking systems, with one model capable of linking 104 languages against 20 million WikiData entities.
- Introduce **Mewslī-9**, a large dataset with nearly 300,000 mentions across 9 diverse languages with links to WikiData. The dataset features many entities that lack English Wikipedia pages and which are thus inaccessible to many prior cross-lingual systems.
- Present frequency-bucketed evaluation that highlights zero- and few-shot challenges with clear headroom, implicitly including low-resource languages without enumerating results over a hundred languages.

2 Task Definition

Multilingual Entity Linking (MEL) is the task of linking an entity mention m in some context language l^c to the corresponding entity $e \in V$ in a *language-agnostic KB*. That is, while the KB may include textual information (names, descriptions, etc.) about each entity in one or more languages, we make no prior assumption about the relationship between these KB languages $L^{\text{kb}} = \{l_1, \dots, l_k\}$ and the mention-side language: l^c may or may not be in L^{kb} .

This is a generalization of *cross-lingual EL* (XEL), which is concerned with the case where $L^{\text{kb}} = \{l'\}$ and $l^c \neq l'$. Commonly, l' is English, and V is moreover limited to the set of entities that express features in l' .

2.1 MEL with WikiData and Wikipedia

As a concrete realization of the proposed task, we use WikiData (Vrandečić and Krötzsch, 2014) as our KB: it covers a large set of diverse entities, is broadly accessible and actively maintained, and it provides access to entity features in many

languages. WikiData itself contains names and short descriptions, but through its close integration with all Wikipedia editions, it also connects entities to rich descriptions (and other features) drawn from the corresponding language-specific Wikipedia pages.

Basing entity representations on features of their Wikipedia pages has been a common approach in EL (e.g. Sil and Florian, 2016; Francis-Landau et al., 2016; Gillick et al., 2019; Wu et al., 2019), but we will need to generalize this to include multiple Wikipedia pages with possibly redundant features in many languages.

2.1.1 WikiData Entity Example

Consider the WikiData Entity *Sí Ràdio*^{Q3511500}, a now defunct Valencian radio station. Its KB entry references Wikipedia pages in three languages, which contain the following descriptions:²

- (Catalan) *Sí Ràdio fou una emissora de ràdio musical, la segona de Radio Autonomía Valenciana, S.A. pertanyent al grup Radiotelevisió Valenciana.*
- (Spanish) *Nou Si Ràdio (anteriormente conocido como Sí Ràdio) fue una cadena de radio de la Comunidad Valenciana y emisora hermana de Nou Ràdio perteneciente al grupo RTVV.*
- (French) *Sí Ràdio est une station de radio publique espagnole appartenant au groupe Ràdio Televisió Valenciana, entreprise de radio-télévision dépendant de la Generalitat valencienne.*

Note that these Wikipedia descriptions are not direct translations, and contain some name variations. We emphasize that this particular entity would have been completely out of scope in the standard cross-lingual task (Tsai and Roth, 2016), because it does not have an English Wikipedia page.

In our analysis, there are millions of WikiData entities with this property, meaning the standard setting skips over the substantial challenges of modeling these (often rarer) entities, and disambiguating them in different language contexts. Our formulation seeks to address this.

²We refer to the first sentence of a Wikipedia page as a description because it follows a standardized format.

2.2 Knowledge Base Scope

Our modeling focus is on using *unstructured textual* information for entity linking, leaving other modalities or structured information as areas for future work. Accordingly, we narrow our KB to the subset of entities that have descriptive text available: We define our entity vocabulary V as all WikiData items that have an associated Wikipedia page in *at least one language*, independent of the languages we actually model.³ This gives 19,666,787 entities, *substantially more than in any other task settings we have found*: the KB accompanying the entrenched TAC-KBP 2010 benchmark (Ji et al., 2010) has less than a million entities, and although English Wikipedia continues to grow, recent work using it as a KB still only contend with roughly 6 million entities (Férvy et al., 2020a; Zhou et al., 2020). Further, by employing a simple rule to determine the set of viable entities, we avoid potential selection bias based on our desired test sets or the language coverage of a specific pretrained model.

2.3 Supervision

We extract a supervision signal for MEL by exploiting the hyperlinks that editors place on Wikipedia pages, taking the anchor text as a linked mention of the target entity. This follows a long line of work in exploiting hyperlinks for EL supervision (Bunescu and Paşca, 2006; Singh et al., 2012; Logan et al., 2019), which we extend here by applying the idea to extract a large-scale dataset of 684 million mentions in 104 languages, linked to WikiData entities. This is at least six times larger than datasets used in prior English-only linking work (Gillick et al., 2019). Such large-scale supervision is beneficial for probing the quality attainable with current-day high-capacity neural models.

3 Mewsli-9 Dataset

We facilitate evaluation on the proposed multilingual EL task by releasing a matching dataset that covers a diverse set of languages and entities.

Mewsli-9 (*Multilingual Entities in News, linked*) contains 289,087 entity mentions appearing in 58,717 originally written news articles from WikiNews, linked to WikiData.⁴

The corpus includes documents in nine languages, representing five language families and

³More details in Appendix C.

⁴www.wikinews.org, using the 2019-01-01 snapshot from archive.org

Lang.	Docs	Mentions	Entities	
			Distinct	∉ EnWiki
ja	3,410	34,463	13,663	3,384
de	13,703	65,592	23,086	3,054
es	10,284	56,716	22,077	1,805
ar	1,468	7,367	2,232	141
sr	15,011	35,669	4,332	269
tr	997	5,811	2,630	157
fa	165	535	385	12
ta	1,000	2,692	1,041	20
en	12,679	80,242	38,697	14
	58,717	289,087	82,162	8,807
en'	1801	2,263	1799	0

Table 1: Corpus statistics for Mewsli-9, an evaluation set we introduce for multilingual entity linking against WikiData. Line en' shows statistics for English WikiNews-2018, by Gillick et al. (2019).

six orthographies.⁵ Per-language statistics appear in Table 1. Crucially, 11% of the 82,162 distinct target entities in Mewsli-9 *do not have English Wikipedia pages*, thereby setting a restrictive upper bound on performance attainable by a standard XEL system focused on English Wikipedia entities.⁶ Even some English documents may contain such mentions, such as the Romanian reality TV show, *Noră pentru mama*^{Q12736895}.

WikiNews articles constitute a somewhat different text genre from our Wikipedia training data: The articles do not begin with a formulaic entity description, for example, and anchor link conventions are likely different. We treat the full dataset as a test set, avoiding any fine-tuning or hyperparameter tuning, thus allowing us to evaluate our model's robustness to domain drift.

Mewsli-9 is a drastically expanded version of the English-only WikiNews-2018 dataset by Gillick et al. (2019). Our automatic extraction technique trades annotation quality for scale and diversity, in contrast to the MEANTIME corpus based on WikiNews (Minard et al., 2016). Mewsli-9 intentionally stretches the KB definition beyond English Wikipedia, unlike VoxEL (Rosales-Méndez et al., 2018). Both MEANTIME and VoxEL are limited to a handful of European languages.

⁵Mewsli-9 languages (code, family, script): Japanese ('ja', Japonic, ideograms); German ('de', Indo-European (IE), Latin); Spanish ('es', IE, Latin); Arabic ('ar', Afro-Asiatic, Arabic); Serbian ('sr', IE, Latin & Cyrillic); Turkish ('tr', Turkic, Latin); Persian ('fa', IE, Perso-Arabic); Tamil ('ta', Dravidian, Brahmic); English ('en', IE, Latin).

⁶As of 2019-10-03.

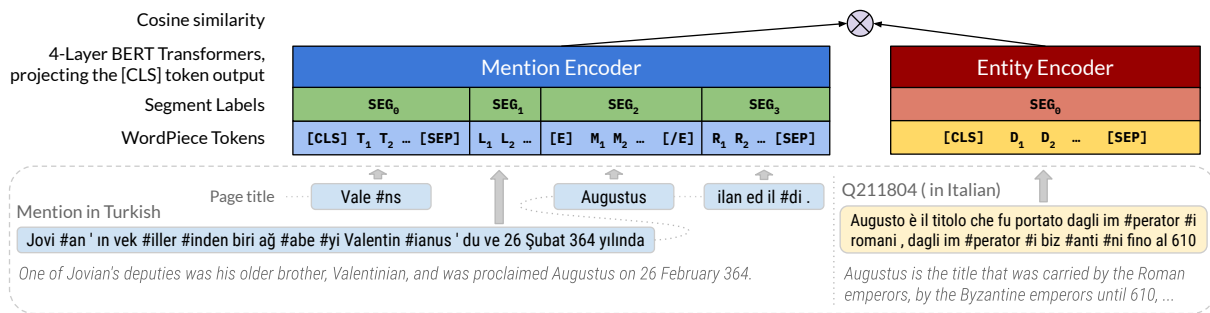


Figure 1: Dual Encoder Model F diagram. The input to the *Mention Encoder* is a sequence of WordPiece tokens that includes the document title (T_i), context immediately left of the mention (L_i), the mention span (M_i) demarcated by [E] and [/E] markers, and context immediately right of the mention (R_i). Segment labels (SEG_i) are also used to distinguish the input segments. The input to the (Model F) *Entity Encoder* is simply the WordPiece tokens in the entity description (D_i). As usual, embeddings passed to the first transformer layer are the sum of positional embeddings (not pictured here), the segment embeddings, and the WordPiece embeddings. The example shows a Turkish mention of Augustus^{Q211804} paired with its Italian description.

4 Model

Prior work showed that a dual encoder architecture can encode entities and contextual mentions in a dense vector space to facilitate efficient entity retrieval via nearest-neighbors search (Gillick et al., 2019; Wu et al., 2019). We take the same approach. The dual encoder maps a mention-entity pair (m, e) to a score:

$$s(m, e) = \frac{\phi(m)^T \psi(e)}{\|\phi(m)\| \|\psi(e)\|}, \quad (1)$$

where ϕ and ψ are learned neural network encoders that encode their arguments as d -dimensional vectors ($d=300$, matching prior work).

Our encoders are BERT-based Transformer networks (Vaswani et al., 2017; Devlin et al., 2019), which we initialize from a pretrained multilingual BERT checkpoint.⁷ For efficiency, we only use the first 4 layers, which results in a negligible drop in performance relative to the full 12-layer stack. The WordPiece vocabulary contains 119,547 symbols covering the top 104 Wikipedia languages by frequency—this is the language set we use in our experiments.

4.1 Mention Encoder

The mention encoder ϕ uses an input representation that is a combination of *local context* (mention span with surrounding words, ignoring sentence boundaries) and simple *global context* (document title). The document title, context, and mention span are marked with special separator tokens as well as identifying token type labels (see Figure 1 for details). Both the mention span markers and

⁷github.com/google-research/bert_multi_cased_L-12_H-768_A-12

document title have been employed in related work (Agarwal and Bikel, 2020; Févry et al., 2020a). We use a maximum sequence length of 64 tokens similar to prior work (Févry et al., 2020a), up to a quarter of which are used for the document title. The CLS token encoding from the final layer is projected to the encoding dimension to form the final mention encoding.

4.2 Entity Encoders

We experiment with two entity encoder architectures. The first, called **Model F**, is a featurized entity encoder that uses a fixed-length text description (64 tokens) to represent each entity (see Figure 1). The same 4-layer Transformer architecture is used—without parameter sharing between mention and entity encoders—and again the CLS token vector is projected down to the encoding dimension. Variants of this entity architecture were employed by Wu et al. (2019) and Logeswaran et al. (2019).

The second architecture, called **Model E** is simply a QID-based embedding lookup as in Févry et al. (2020a). This latter model is intended as a baseline. *A priori*, we expect Model E to work well for common entities, less well for rarer entities, and not at all for zero-shot retrieval. We expect Model F to provide more parameter-efficient storage of entity information and possibly improve on zero- and few-shot retrieval.

4.2.1 Entity Description Choice

There are many conceivable ways to make use of entity descriptions from multiple languages. We limit the scope to using one primary description per entity, thus obtaining a single coherent text fragment to feed into the Model F encoder.

We use a simple data-driven selection heuristic that is based on observed entity usage: Given an entity e , let $n_e(l)$ denote the number of mentions of e in documents of language l , and $n(l)$ the global number of mentions in language l across all entities. From a given source of descriptions—first Wikipedia and then WikiData—we order the candidate descriptions $(t_e^{l_1}, t_e^{l_2}, \dots)$ for e first by the per-entity distribution $n_e(l)$ and then by the global distribution $n(l)$.⁸ For the example entity in Section 2.1.1, this heuristic selects the Catalan description because 9/16 training examples link to the Catalan Wikipedia page.

4.3 Training Process

In all our experiments, we use an 8k batch size with in-batch sampled softmax (Gillick et al., 2018). Models are trained with Tensorflow (Abadi et al., 2016) using the Adam optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). All BERT-based encoders are initialized from a pretrained checkpoint, but the Model E embeddings are initialized randomly. We doubled the batch size until no further held-out set gains were evident and chose the number of training steps to keep the training time of each phase under one day on a TPU. Further training would likely yield small improvements. See Appendix B for more detail.

5 Experiments

We conduct a series of experiments to gain insight into the behavior of the dual encoder retrieval models under the proposed MEL setting, asking:

- What are the relative merits of the two types of entity representations used in Model E and Model F (embeddings vs. encodings of textual descriptions)?
- Can we adapt the training task and hard-negative mining to improve results across the entity frequency distribution?
- Can a single model achieve reasonable performance on over 100 languages while retrieving from a 20 million entity candidate set?

5.1 Evaluation Data

We follow Upadhyay et al. (2018) and evaluate on the “hard” subset of the Wikipedia-derived test

⁸The candidate descriptions (but not V) are limited to the 104 languages covered by our model vocabulary—in general, both Wikipedia and WikiData cover more than 300 languages.

set introduced by Tsai and Roth (2016) for cross-lingual EL against English Wikipedia, **TR2016^{hard}**. This subset comprises mentions for which the correct entity did not appear as the top-ranked item in their alias table, thus stress-testing a model’s ability to generalize beyond mention surface forms.

Unifying this dataset with our task formulation and data version requires mapping its gold entities from the provided, older Wikipedia titles to newer WikiData entity identifiers (and following intermediate Wikipedia redirection links). This succeeded for all but 233/42,073 queries in TR2016^{hard}—our model receives no credit on the missing ones.

To be compatible with the pre-existing train/test split, we excluded from our training set all mentions appearing on Wikipedia pages in the full TR2016 test set. This was done for all 104 languages, to avoid cross-lingual overlap between train and test sets. This aggressive scheme holds out 33,460,824 instances, leaving our final training set with 650,975,498 mention-entity pairs. Figure 2 provides a break-down by language.

5.2 Evaluating Design Choices

5.2.1 Setup and Metrics

In this first phase of experiments we evaluate design choices by reporting the *differences* in Recall@100 between two models at a time, for conciseness. Note that for final system comparisons, it is standard to use Accuracy of the top retrieved entity (R@1), but to evaluate a dual encoder retrieval model, we prefer R@100 as this is better matched to its likely use case as a candidate generator.

Here we use the TR2016^{hard} dataset, as well as a portion of the 104-language set held out from our training data, sampled to have 1,000 test mentions per language. (We reserve the new Mewsl-9 dataset for testing the final model in Section 5.5.)

Reporting results for 104 languages is a challenge. To break down evaluation results by entity frequency bins, we partition a test set according to the frequency of its gold entities as observed in the training set. This is in line with recent recommendations for finer-grained evaluation in EL (Waitelonis et al., 2016; Ilievski et al., 2018).

We calculate metrics within each bin, and report macro-average over bins. This is a stricter form of the label-based macro-averaging sometimes used, but better highlights the zero-shot and few-shot cases. We also report micro-average metrics, computed over the entire dataset, without binning.

Bin	(a)		(b)		(c)	
	holdout	TR2016 ^{hard}	holdout	TR2016 ^{hard}	holdout	TR2016 ^{hard}
[0, 1)	+0.842	+0.380	+0.009	+0.093	+0.044	+0.144
[1, 10)	+0.857	+0.814	+0.018	+0.037	+0.051	+0.031
[10, 100)	+0.211	+0.191	+0.012	+0.024	+0.006	-0.019
[100, 1k)	-0.010	-0.031	+0.007	+0.019	-0.005	-0.015
[1k, 10k)	-0.018	-0.051	+0.008	+0.011	-0.003	-0.007
[10k, +)	-0.009	-0.089	+0.004	+0.003	-0.002	-0.013
micro-avg	+0.018	+0.008	+0.006	+0.017	-0.001	-0.006
macro-avg	+0.312	+0.202	+0.010	+0.031	+0.015	+0.020

Table 2: R@100 differences between pairs of models: (a) model F (featurized inputs for entities) relative to model E (dedicated embedding for each entity); (b) add cross-lingual entity-entity task on top of the mention-entity task for model F; (c) control label balance per-entity during negative mining (versus not).

5.2.2 Entity Encoder Comparison

We first consider the choice of entity encoder, comparing Model F with respect to Model E.

Table 2(a) shows that using the entity descriptions as inputs leads to dramatically better performance on rare and unseen entities, in exchange for small losses on entities appearing more than 100 times, and overall improvements in both macro and micro recall.

Note that as expected, the embedding Model E gives 0% recall in zero-shot cases, as their embeddings are randomly initialized and never get updated in absence of any training examples.

The embedding table of Model E has 6 billion parameters, but there is no sharing across entities. Model F has approximately 50 times fewer parameters, but can distribute information in its shared, compact WordPiece vocabulary and Transformer layer parameters. We can think of these dual encoder models as classifiers over 20 million classes where the softmax layer is either parameterized by an ID embedding (Model E) or an encoding of a description of the class itself (Model F). Remarkably, using a Transformer for the latter approach effectively compresses (nearly) all the information in the traditional embedding model into a compact and far more generalizable model.

This result highlights the value of analyzing model behavior in terms of entity frequency. When looking at the micro-averaged metric in isolation, one might conclude that the two models perform similarly; but the macro-average is sensitive to the large differences in the low-frequency bins.

5.2.3 Auxiliary Cross-Lingual Task

In seeking to improve the performance of Model F on tail entities, we return to the (partly redundant) entity descriptions in multiple languages. By choosing just one language as the input, we are ignoring potentially valuable information in the remaining descriptions.

Here we add an auxiliary task: cross-lingual entity description retrieval. This reuses the entity encoder ψ of Model F to map two descriptions of an entity e to a score, $s(t_e^l, t_e^{l'}) \propto \psi(t_e^l)^T \psi(t_e^{l'})$, where $t_e^{l'}$ is the description selected by the earlier heuristic, and t_e^l is sampled from the other available descriptions for the entity.

We sample up to 5 such cross-lingual pairs per entity to construct the training set for this auxiliary task. This makes richer use of the available multilingual descriptions, and exposes the model to 39 million additional high-quality training examples whose distribution is decoupled from that of the mention-entity pairs in the primary task. The multi-task training computes an overall loss by averaging the in-batch sampled softmax loss for a batch of (m, e) pairs and for a batch of (e, e) pairs.

Table 2(b) confirms this brings consistent quality gains across all frequency bins, and more so for uncommon entities. Again, reliance on the micro-average metric alone understates the benefit in this data augmentation step for rarer entities.

5.2.4 Hard-Negative Mining

Training with hard-negatives is highly effective in monolingual entity retrieval (Gillick et al., 2019), and we apply the technique they detail to our multilingual setting.

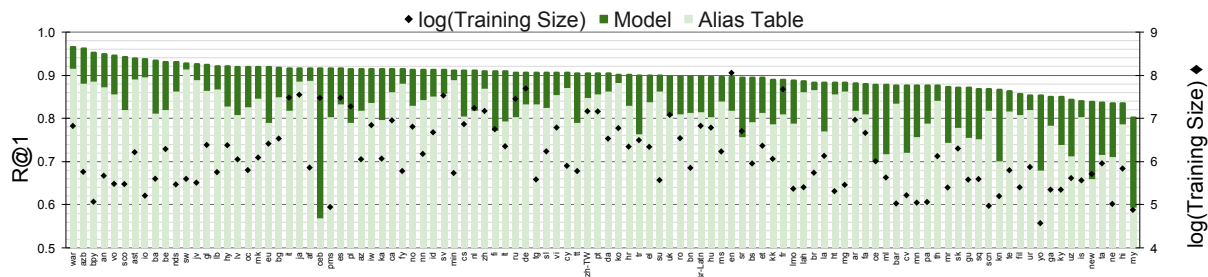


Figure 2: Accuracy of Model F⁺ on the 104 languages in our balanced Wikipedia heldout set, overlaid on alias table accuracy and Wikipedia training set size. (See Figure B1 in the Appendix for a larger view.)

In its standard form, a certain number of negatives are mined for each mention in the training set by collecting top-ranked but incorrect entities retrieved by a prior model. However, this process can lead to a form of the class imbalance problem as uncommon entities become over-represented as negatives in the resulting data set. For example, an entity appearing just once in the original training set could appear hundreds or thousands of times as a negative example. Instead, we control the ratio of positives to negatives on a per-entity basis, mining up to 10 negatives per positive.

Table 2(c) confirms that our strategy effectively addresses the imbalance issue for rare entities with only small degradation for more common entities. We use this model to perform a second, final round of the adapted negative mining followed by further training to improve on the macro-average further by +.05 (holdout) and +.08 (TR2016^{hard}).

The model we use in the remainder of the experiments combines all these findings. We use Model F with the entity-entity auxiliary task and hard negative mining with per-entity label balancing, referenced as **Model F⁺**.

5.3 Linking in 100 Languages

Breaking down the model’s performance by language (R@1 on our heldout set) reveals relatively strong performance across all languages, despite greatly varying training sizes (Figure 2). It also shows improvement over an alias table baseline on all languages. While this does not capture the relative difficulty of the EL task in each language, it does strongly suggest effective cross-lingual transfer in our model: even the most data-poor languages have reasonable results. This validates our massively multilingual approach.

	Tsai+	Upad.+	Model F ⁺
Languages	13	5	104
 V 	5m	5m	20m
Candidates	20	20	20m
de	0.53	0.55	0.62
es	0.54	0.57	0.58
fr	0.48	0.51	0.54
it	0.48	0.52	0.56
Average	0.51	0.54	0.57

Table 3: Our best model outperforms previous related non-monolingual models that relied on alias tables and disambiguated among a much smaller set of entities. *Bottom half*: linking accuracy on the TR2016^{hard} test set. *Top half*: language coverage; entity vocabulary size; and entities disambiguated among at inference time. *Middle columns*: (Tsai and Roth, 2016) and (Upadhyay et al., 2018).

5.4 Comparison to Prior Work

We evaluate the performance of our final retrieval model relative to previous work on two existing datasets, noting that direct comparison is impossible because our task setting is novel.

5.4.1 Cross-Lingual Wikification Setting

We compare to two previously reported results on TR2016^{hard}: the WIKIME model of Tsai and Roth (2016) that accompanied the dataset, and the XELMS-MULTI model by Upadhyay et al. (2018). Both models depend at their core on multilingual word embeddings, which are obtained by applying (bilingual) alignment or projection techniques to pretrained monolingual word embeddings.

As reported in Table 3, our multilingual dual encoder outperforms the other two by a significant margin. To the best of our knowledge, this is the highest accuracy to-date on this challenging evaluation set. (Our comparison is limited to the four languages on which Upadhyay et al. (2018) evaluated their multilingual model.)

	DEER	Model F ⁺
Languages	1	104
Candidates = V	5.7m	20m
R@1	0.92	0.92
R@100	0.98	0.99

Table 4: Comparison to DEER model (Gillick et al., 2019) on their English WikiNews-2018 dataset.

This is a strong validation of the proposed approach because the experimental setting is heavily skewed toward the prior models: Both are *rerankers*, and require a first-stage candidate generation step. They therefore only disambiguate among the resulting ≤ 20 candidate entities (only from English Wikipedia), whereas our model performs retrieval against all 20 million entities.

5.4.2 Out-of-Domain English Evaluation

We now turn to the question of how well the proposed multilingual model can maintain competitive performance in English and generalize to a domain other than Wikipedia. Gillick et al. (2019) provides a suitable comparison point. Their DEER model is closely related to our approach, but used a more light-weight dual encoder architecture with bags-of-embeddings and feed-forward layers without attention and was evaluated on English EL only. On the English WikiNews-2018 dataset they introduced, our Transformer-based multilingual dual encoder matches their monolingual model’s performance at R@1 and improves R@100 by 0.01 (reaching 0.99) Our model thus retains strong English performance despite covering many languages and linking against a larger KB. See Table 4.

5.5 Evaluation on Mewsli-9

Table 5 shows the performance of our model on our new Mewsli-9 dataset compared with an alias table baseline that retrieves entities based on the prior probability of an entity given the observed mention string. Table 6 shows the usual frequency-binned evaluation. While overall (micro-average) performance is strong, there is plenty of headroom in zero- and few-shot retrieval.

5.5.1 Example Outputs

We sampled the model’s correct predictions on Mewsli-9, focusing on cross-lingual examples where entities do not have an English Wikipedia page (Table 7). These examples demonstrate that

Language	Alias Table		Model F ⁺	
	R@1	R@10	R@1	R@10
ar	0.89	0.93	0.92	0.98
de	0.86	0.91	0.92	0.97
en	0.79	0.86	0.87	0.94
es	0.82	0.90	0.89	0.97
fa	0.87	0.92	0.92	0.97
ja	0.82	0.90	0.88	0.96
sr	0.87	0.92	0.93	0.98
ta	0.79	0.85	0.88	0.97
tr	0.80	0.88	0.88	0.97
micro-avg	0.83	0.89	0.89	0.96
macro-avg	0.83	0.89	0.90	0.97

Table 5: Results of our main dual encoder Model F⁺ on the new Mewsli-9 dataset. Consistent performance across languages in a different domain from the training set points at good generalization.

Bin	Queries	Model F ⁺		+CA
		R@1	R@10	R@1
[0, 1)	3,198	0.08	0.34	0.07
[1, 10)	6,564	0.58	0.81	0.60
[10, 100)	32,371	0.80	0.93	0.82
[100, 1k)	66,232	0.90	0.97	0.90
[1k, 10k)	78,519	0.93	0.98	0.93
[10k, +)	102,203	0.94	0.99	0.96
micro-avg	289,087	0.89	0.96	0.91
macro-avg		0.70	0.84	0.71

Table 6: Results on the new Mewsli-9 dataset, by entity frequency, attained by our main dual encoder Model F⁺, plus reranking its predictions with a Cross-Attention scoring model (CA).

the model effectively learns cross-lingual entity representations. Based on a random sample of the model’s errors, we also show examples that summarize notable error categories.

5.5.2 Reranking Experiment

We finally report a preliminary experiment to apply a cross-attention scoring model (CA) to rerank entity candidates retrieved by the main dual encoder (DE), using the same architecture of Logeswaran et al. (2019). We feed the concatenated mention text and entity description into a 12-layer Transformer model, initialized from the same multilingual BERT checkpoint referenced earlier.

The CA model’s CLS token encoding is used to classify mention-entity coherence. We train the model with a binary cross-entropy loss, using posi-

Context 1	... Bei den neuen Bahnen handelt es sich um das Model Tramino von der polnischen Firma Solaris Bus & Coach...
Prediction	Solaris Tramino ^{Q780281} : Solaris Tramino – rodzina tramwajów, które są produkowane przez firmę Solaris Bus & Coach z Bolechowa koło Poznania...
Outcome	<i>Correct: A family of trams originally manufactured in Poland, mentioned here in German, linked to its Polish description.</i>
Context 2	... sobre una tecnología que permitiría fabricar chocolate a partir de los zumos de fruta , agua con vitamina C o gaseosa dietética...
Prediction	fruit juice ^{Q20932605} : Fruchtsaft, spezieller auch Obstsaft, ist ein aus Früchten einer oder mehrerer Fruchtarten gewonnenes flüssiges Erzeugnis...
Outcome	<i>Correct: A Spanish mention of “fruit juice” linked to its German description—only “juice” has a dedicated English Wikipedia page.</i>
Context 3	... Душан Ивковић рекао је да је његов тим имао императив победе над (Италијом) на Европском првенству...
Prediction	It. men’s water polo team ^{Q261190} : La nazionale di pallanuoto maschile dell’ Italia...
Expected	It. nat. basketball team ^{Q261190} : La nazionale di pallacanestro italiana è la selezione dei migliori giocatori di nazionalità italiana...
Outcome	<i>Wrong: A legitimately ambiguous mention of “Italy” in Serbian (sports context), for which model retrieved the water polo and football teams, followed by the expected basketball team entity, all featurized in Italian.</i>
Context 4	... In July 2009, action by the Federal Bureau of Reclamation to protect threatened fish stopped irrigation pumping to parts of the California Central Valley...
Prediction	irrigation sprinkler ^{Q998539} : スプリンクラーは、水に高圧をかけ飛沫にしてノズルから散布する装置
Outcome	<i>Wrong: Metonymous mention of Central Valley Project^{Q2944429} in English, but model retrieved the more literal match, featurized in Chinese. Metonymy is a known challenging case for EL (Ling et al., 2015).</i>

Table 7: Correct and mistaken examples observed in error analysis of dual encoder model F⁺ on Mewsl-9.

tives from our Wikipedia training data, taking for each one the top-4 DE-retrieved candidates plus 4 random candidates (proportional to the positive distributions).

We use the trained CA model to rerank the top-5 DE candidates for Mewsl-9 (Table 6). We observed improvements on most frequency buckets compared to DE R@1, which suggests that the model’s few-shot capability can be improved by cross-lingual reading-comprehension. This also offers an initial multilingual validation of a similar two-step BERT-based approach recently introduced in a monolingual setting by (Wu et al., 2019), and provides a strong baseline for future work.

6 Conclusion

We have proposed a new formulation for multilingual entity linking that seeks to expand the scope of entity linking to better reflect the real-world challenges of rare entities and/or low resource languages. Operationalized through Wikipedia and WikiData, our experiments using enhanced dual encoder retrieval models and frequency-based evaluation provide compelling evidence that it is feasible

to perform this task with a single model covering over a 100 languages.

Our automatically extracted Mewsl-9 dataset serves as a starting point for evaluating entity linking beyond the entrenched English benchmarks and under the expanded multilingual setting. Future work could investigate the use of non-expert human raters to improve the dataset quality further.

In pursuit of improved entity representations, future work could explore the joint use of complementary multi-language descriptions per entity, methods to update representations in a light-weight fashion when descriptions change, and incorporate relational information stored in the KB.

Acknowledgments

Thanks to Sayali Kulkarni and the anonymous reviewers for their helpful feedback.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale

- machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.
- Oshin Agarwal and Daniel M Bikel. 2020. Entity linking via dual and cross-attention encoders. *arXiv preprint arXiv:2004.03555*.
- Razvan Bunescu and Marius Paşca. 2006. [Using encyclopedic knowledge for named entity disambiguation](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Thibault Févry, Nicholas FitzGerald, Livio Baldini Soares, and Tom Kwiatkowski. 2020a. [Empirical evaluation of pretraining strategies for supervised entity linking](#). In *Automated Knowledge Base Construction*.
- Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020b. Entities as experts: Sparse memory access with entity supervision. *arXiv preprint arXiv:2004.07202*.
- Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. [Capturing semantic similarity for entity linking with convolutional neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California. Association for Computational Linguistics.
- Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. [Learning dense representations for entity retrieval](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.
- Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. [End-to-end retrieval in continuous space](#). *arXiv preprint arXiv:1811.08008*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papatat, and Ming-Wei Chang. 2020. [REALM: Retrieval-augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria. PMLR.
- Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. [Systematic study of long tail phenomena in entity linking](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Grifflitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proceedings of the Text Analytics Conference*.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *International Conference on Learning Representations*, San Diego, CA.
- Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. [Design challenges for entity linking](#). *Transactions of the Association for Computational Linguistics*, 3:315–328.
- Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. [Barack’s wife Hillary: Using knowledge graphs for fact-aware language modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. [Zero-shot entity linking by reading entity descriptions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. [Cross-language entity linking](#). In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.
- Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. [MEANTIME, the](#)

- NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).
- 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)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. [Data-to-text generation with entity modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035. 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.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.
- Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. VoxEL: a benchmark dataset for multilingual entity linking. In *International Semantic Web Conference*, pages 170–186. Springer.
- Peter Shaw, Philip Massey, Angelica Chen, Francesco Piccinno, and Yasemin Altun. 2019. [Generating logical forms from graph representations of text and entities](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 95–106. Association for Computational Linguistics.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.
- Avirup Sil and Radu Florian. 2016. [One for all: Towards language independent named entity linking](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2255–2264, Berlin, Germany. Association for Computational Linguistics.
- Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.
- Chen-Tse Tsai and Dan Roth. 2016. [Cross-lingual wikification using multilingual embeddings](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.
- Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. [Joint multilingual supervision for cross-lingual entity linking](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Jörg Waitelonis, Henrik Jürges, and Harald Sack. 2016. [Don’t compare apples to oranges: Extending gerbil for a fine grained nel evaluation](#). In *Proceedings of the 12th International Conference on Semantic Systems, SEMANTiCS 2016*, page 65–72, New York, NY, USA. Association for Computing Machinery.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.
- Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. [Towards zero-resource cross-lingual entity linking](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 243–252. Association for Computational Linguistics.
- Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. [Improving candidate generation for low-resource cross-lingual entity linking](#). *Transactions of the Association for Computational Linguistics*, 8:109–124.

Appendix

A Mewsli-9 Dataset

Available at: <http://goo.gle/mewsli-dataset>

We used an automated process to construct Mewsli-9, exploiting link anchor text to identify naturally occurring entity mentions in WikiNews articles, from its inception to the end of 2018.

From a given WikiNews page dump,⁹ we extracted text including link anchors and section headings using a modified version of `wikiextractor`.¹⁰

To obtain clean article text, we discard page-final sections that merely contain external references, etc. This is done by matching section headings against a small set of hand-collected, language-specific patterns.

Mention candidates are filtered to those remaining links that point to Wikipedia pages in any language (not limited to our 104 languages). These Wikipedia links are redirected if necessary, and resolved to WikiData identifiers to determine the gold entity for a mention. There are many reasons why resolution may fail, including mistakes in the original markup and churn in the data sources over time. The final dataset is limited to (mention, entity) pairs where resolution to WikiData succeeded.

B Training Details and Hyperparameters

All model training was carried out on a Google TPU v3 architecture,¹¹ using batch size 8192 and a learning rate schedule that uses linear warm-up followed by linear decay to 0.

The first phase of training our dual encoders (DE) with in-batch random negatives encompasses 500,000 steps, which takes approximately one day.

Where hard-negative training is applied, we initialize from the corresponding prior model checkpoint and continue training against the multi-task loss for a further 250,000 steps, which also takes about a day.

Other than the limit to using a 4-layer Transformer stack, our mention encoder and model F entity encoders use the same hyperparameters as mBERT-base, allowing initialization from the publicly available checkpoint—we use the weights of its first 4 layers, in addition to those of the token and positional embeddings.

⁹archive.org/download/XXwikinews-20190101/XXwikinews-20190101-pages-articles.xml.bz2 where XX is a language code.

¹⁰github.com/attardi/wikiextractor

¹¹cloud.google.com/tpu/docs/tpus

QID	label
Q4167836	category
Q24046192	category stub
Q20010800	user category
Q11266439	template
Q11753321	navigational template
Q19842659	user template
Q21528878	redirect page
Q17362920	duplicated page
Q14204246	project page
Q21025364	project page
Q17442446	internal item
Q26267864	KML file
Q4663903	portal
Q15184295	module

Table 8: WikiData identifiers used for filtering out Wikimedia-internal entities from our KB definition.

The cross-attention scoring model (CA) in the final preliminary experiment is a full 12-layer Transformer (also mBERT-base), and was trained for 1 million steps, taking just under one day.

The learning rates were 1e-4 (DE) and 1-e5 (CA) and included warm-up phases of 10% (DE) and 1% (CA) of the respective number of training steps.

C Data Preprocessing

We used the 2019-10-03 dump of Wikipedia and WikiData, parsed using in-house tools.¹²

Two filtering criteria are relevant in preprocessing WikiData to define our KB. The first is to exclude items that are a subclass (P279) or instance of (P31) the most common Wikimedia-internal administrative entities, detailed in Table 8. The remaining entities are then filtered to retain only those for which the WikiData entry points to at least one Wikipedia page, in any language, motivated by our objective of using descriptive text as entity features.

D Other data sources

– TR2016^{hard} dataset: cogcomp.seas.upenn.edu/page/resource_view/102

– English WikiNews-2018 dataset by (Gillick et al., 2019): github.com/google-research/google-research/tree/master/dense_representations_for_entity_retrieval

¹²dumps.wikimedia.org

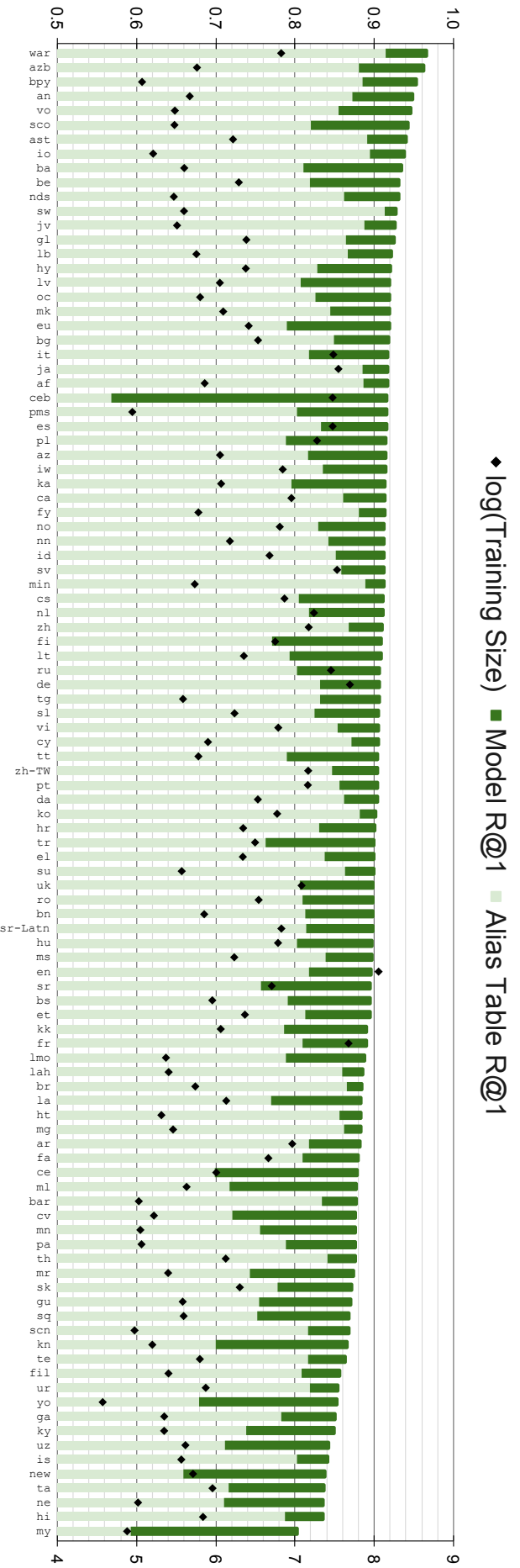


Figure B1: Accuracy (left-axis) of Model F⁺ on the 104 languages in our balanced Wikipedia heldout set, overlaid on alias table accuracy, and Wikipedia training set size (right-axis). Our Model F⁺ obtains relatively strong performance across 104 languages and outperforms an alias table baseline in all cases. Although task difficulty is not necessarily comparable between languages, this result suggests effective cross-lingual transfer is happening: even languages with small training sets show reasonable accuracy. (*This is a scaled-up reproduction of Figure 2 from Section 5.3.*)