When your Cousin has the Right Connections: Unsupervised Bilingual Lexicon Induction for Related Data-Imbalanced Languages

Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, Rachel Bawden

When your Cousin has the Right Connections: Unsupervised Bilingual Lexicon Induction for Related Data-Imbalanced Languages

Niyati Bafna,[1] **Cristina España-Bonet,**[3] **Josef van Genabith,**[2,3]
Benoît Sagot,[1] **Rachel Bawden**[1]
[1]Inria, Paris, France
[2]Saarland University, Saarland Informatics Campus, Germany
[3]DFKI GmbH, Saarland Informatics Campus, Germany
nbafna1@jhu.edu, {josef.van_genabith, cristinae}@dfki.de
{benoit.sagot, rachel.bawden}@inria.fr

Abstract

Most existing approaches for unsupervised bilingual lexicon induction (BLI) depend on good quality static or contextual embeddings requiring large monolingual corpora for both languages. However, unsupervised BLI is most likely to be useful for low-resource languages (LRLs), where large datasets are not available. Often we are interested in building bilingual resources for LRLs against related high-resource languages (HRLs), resulting in severely imbalanced data settings for BLI. We first show that state-of-the-art BLI methods in the literature exhibit near-zero performance for severely data-imbalanced language pairs, indicating that these settings require more robust techniques. We then present a new method for unsupervised BLI between a related LRL and HRL that only requires inference on a masked language model of the HRL, and demonstrate its effectiveness on truly low-resource languages Bhojpuri and Magahi (with <5M monolingual tokens each), against Hindi. We further present experiments on (mid-resource) Marathi and Nepali to compare approach performances by resource range, and release our resulting lexicons for five low-resource Indic languages: Bhojpuri, Magahi, Awadhi, Braj, and Maithili, against Hindi.

Keywords: bilingual lexicon induction, low-resource, Indic languages

1. Introduction

Bilingual lexicons are a basic resource with varied uses, both in themselves, for dictionary building and language learning, as well as seeds for solving other problems in natural language processing (NLP), such as parsing (Zhao et al., 2009; Durrett et al., 2012) and word-to-word and unsupervised machine translation (Irvine and Callison-Burch, 2013; Thompson et al., 2019).

While there is growing interest in unsupervised or minimally supervised bilingual lexicon induction (BLI), existing methods often depend on aligning monolingual word embedding spaces, assumed to be of good quality for both languages, and/or bilingual supervision (Artetxe et al., 2016, 2017; Conneau et al., 2018; Artetxe et al., 2018a,b, 2019). However, extremely low-resource languages (LRLs) and dialects often lack good quality embeddings due to limited monolingual data, leading to very low or near-zero performance of alignment-based methods for these languages (Wada et al., 2019; Eder et al., 2021).

This is the case for the under-researched Indic language continuum, which is the focus of this article (see Section 2 for a description of the linguistic setup in India that motivates our work). We work with five extremely low-resourced Indic languages, Bhojpuri (bho), Magahi mag), Awadhi (awa), Maithili (mai), and Braj (bra), which are closely related to higher-resourced Hindi, and which have extremely limited resources, in terms of training data (<5M tokens of monolingual data) and embeddings, and even evaluation data. We demonstrate that state-of-the-art, alignment-based methods perform poorly in these settings, and introduce a new method for unsupervised BLI that performs much better. We aim to design methods that work well in characteristic data-scarce conditions, as well as generate resources for further work in these languages.

Our main contribution is a novel unsupervised BLI method to address the typical scenario of the LRLs of the Indic continuum, i.e. for extremely LRLs that share significant overlap with a closely related HRL. We suppose that a masked language model (MLM) such as monolingual BERT (Devlin et al., 2019) is available for the HRL and that we have some monolingual LRL sentences that contain unknown words. The method consists of building a lexicon iteratively by using the HRL MLM over LRL sentences to extract translation equivalents, and replacing learnt words in LRL sentences with HRL equivalents to make them more tractable for the HRL MLM for future unknown words (see Section 4).

Given the lack of existing gold lexicons for our target languages (a frequent scenario for extremely LRLs), we create silver lexicons for Bhojpuri and Magahi created from parallel data, unfortunately unavailable for Awadhi, Maithili, and Braj. We also perform control experiments on Marathi and Nepali,

two medium-resource languages more distantly related to Hindi with available gold lexicons, and discuss the performance of canonical methods and our proposed method on these languages, shedding light on what strategies are appropriate for differently-resourced language pairs. Our experiments indicate that current state-of-the-art methods are not suitable for low-resourced dialects, and methods that account for the data imbalance in the language pair, such as ours, may be more successful. We release our code, our generated lexicons for all languages (to our knowledge the first to be publicly released for all languages except Bhojpuri),[1] and our created silver evaluation lexicons for Bhojpuri and Magahi.[2] See details of our released lexicons in Section 7.

Our motivation and method, while relevant to the 40+ resource-scarce languages of the Indic language family and other Indian languages, are also relevant to other linguistic systems with similar circumstances, i.e. with a single high-or-medium resource language (usually a standard dialect), and several closely related dialects with lexical, morphological, and syntactic variation, written in the same script with or without orthographic standardization. This setup describes, for example, the Arabic continuum, the Turkic language continuum, and the German dialect system.

2. Linguistic Setup in India

India has around 15-22 languages that are medium-to-high-resource, such as Hindi, Marathi, and Tamil, but dozens of other languages and dialects that are extremely low-resourced, with very little monolingual data (<5M tokens), and no other resources, such as Marwadi, Tulu, Dogri, and Santhali. These languages are often closely related to at least one high-resource language (HRL), meaning that they share morphosyntactic properties as well as a high number of cognates (Jha, 2019; Mundotiya et al., 2021) (see Table 1 for examples). They often have no official status in the regions where they are spoken, and therefore do not have concerted funding efforts for data collection or research. Even when such efforts do exist,[3] the collected corpora are rarely of the magnitude at which static or contextual embeddings can be well-estimated. While the actual number of distinct dialects and languages spoken in India is contested, people self-reported about 576 such "mother tongue" dialects in the latest census,[4] which were then grouped into around 121 languages. Only 22 of these languages have official status (i.e. they are either the official language of some state/union territory, or have national cultural significance), and are therefore accorded funds for the development of resources.

Therefore, although some studies in the literature question the real use case for entirely unsupervised BLI (Vulić et al., 2019), since it is "easy" to collect a small bilingual lexicon, we argue that situations such as these, where there is a large number of languages to build support for, and where efforts in data collection and annotation for individual languages are restricted by the availability of funds, do constitute genuine application scenarios for unsupervised BLI.

Furthermore, we focus on a scenario where the two languages in question are closely related. This is because for most of the low-resource languages in the Indian context cited above, we can usually find a linguistic neighbour that is relatively well off, usually one of India's 22 scheduled languages.[5] In general, when building resources for a given low-resource dialect or language, it is likely that the standard variant of that dialect, or the HRL closest to it, will have large enough corpora available to build a good quality MLM. We target our efforts to these situations.

3. Related Work

Recent years have seen interest in unsupervised BLI (Haghighi et al., 2008; Artetxe et al., 2016, 2017; Conneau et al., 2018; Artetxe et al., 2018a,b, 2019), allowing the possibility of BLI for LRLs. Most unsupervised approaches, notably MUSE (Conneau et al., 2018) and VecMap (Artetxe et al., 2016, 2017, 2018a,b) are based on training static embeddings from large monolingual corpora (Mikolov et al., 2013; Bojanowski et al., 2017), and aligning the embeddings using linear or non-linear mappings, using an initial seed (Xing et al., 2015; Artetxe et al., 2016).

Recent works have also looked at using contextual embeddings or BERT-based models (Peters et al., 2018; Devlin et al., 2019; Ruder et al., 2019) for BLI. Gonen et al. (2020) induce word-level trans-

[1] Although dedicated teams are working towards building resources for these languages (Mundotiya et al., 2021), these resources (including evaluation resources) have not yet been made public as far as we know.

[2] Code and resources available here: https://github.com/niyatibafna/BLI-for-Indic-languages.

[3] See https://data.ldcil.org/text.

[4] See https://www.outlookindia.com/national/explained-what-is-mother-tongue-survey-and-its-importance-in-preserving-india-s-linguistic-data-news-235854.

[5] Of course, there are exceptions to this observation; naturally, this is not true for language isolates such as Burushaski, spoken in Pakistan, or for the small minority of Austro-Asiatic langauges such as Mundari, spoken in India, which do not have a single (Indian) high-resource sister language.

Meaning	boy (nom)	sister (nom)	your (hon., fem. sing. obj)	told (completive)	(you) are going
Hindi	ləɖkɑː	bəhən	aːpkiː	bət̪ːaːjaː/ kəːh lijaː	dʑaː rəheː hoː
Awadi	ləɖkɑː	bəhin	aːpən	bət̪ːaːvət̪	dʑaːt̪ əhaːi
Bhojpuri	ləikaː	bəhin	aːpən	kəhəl	dʑaːt̪ baː
Magahi	ləiːkaː	bəhin	əpən	kəhəlieː	dʑaː həi
Maithili	ləɖkɑː	bəhin	əhaːnk	kəhəlhuⁿ	dʑaː rəhəl əʧʰ i

Table 1: Examples of cognates. Since the Devanagari script is phonetically transparent, phonetic similarity is visible both in IPA and in Devanagari (not shown).

Input and Output Examples for Bhojpuri		
1	Input	उल्लास और अध्यात्मिका से [MASK] आपके तीर्थ यात्रा आनंदमय हो। joy and spirituality-with [MASK] your pilgrimage enjoyable may-be 'May your pilgrimage be filled with joy and spirituality.'
	Mask	भरल 'filled'
	Correct	भरी
	Preds	परिपूर्ण, भरी, युक्त, भरपूर, सम्पन्न replete, filled, containing, filled-up, prosperous
2	Input	प्रधानमंत्री सम्मेलन में भईल विचार-विमर्श अउर इनपुट बतवला के तारीफ [MASK] । Prime Minister conference in occurred discussion and input telling-of praise [MASK] . 'The Prime Minister praised the discussion and inputs made in the conference.'
	Mask	कइलन 'did'
	Correct	की, करी
	Preds	करे, करी, की, किया, *करेल do-hypothetical, did-fem, did-fem, did-masc, -
3	Input	हमनी के **उ** [MASK] पर बहुते गर्व बा । I/We those [MASK] on lots of pride was . 'I/We was/were very proud of those people.'
	Mask	लोगन 'people'
	Correct	लोग, लोगों
	Preds	बात, काम, लड़की, दिन, औरत thing, work, girl, day, woman
	New input	हमनी के **उन** [MASK] पर बहुते गर्व बा ।
	Preds	सब, लोग, लोगों, दिन, सभी all of (them), people, people, day, all of (them)

Table 2: I/O examples, shown for Bhojpuri and the outputs of Basic. **Input**: Target language text with an unknown masked word. **Mask**: the original masked target language word. **Correct**: Acceptable Hindi equivalents given in the silver lexicon. **Preds**: The top predictions made by the mask-filling model. We choose 5 representative examples here(in practice we use the top 30). The underlined prediction is the one that has the lowest normalized edit distance with the original target language word, and would therefore be chosen in the Basic approach as the top candidate. Blue highlighted predictions are correct equivalents, even if not included in silver lexicon. Orange highlighted examples fit the mask but are not equivalents of the masked target word. * indicates a non-word in Hindi. **New input**: Input with known source word replaced with target equivalent (indicated in bold).

lations by directly prompting mBERT (Devlin et al., 2019). Yuan et al. (2020) present a human-in-the-loop system for BLI in four low-resource languages, updating contextual embeddings with the help of annotations provided by a native speaker. Zhang et al. (2021) present CSCBLI, a method that uses a "spring network" to align non-isomorphic contextual embeddings, and interpolates them with static embeddings to estimate word similarities, showing superior results to other methods using contextual embeddings, notably BLISS (Patra et al., 2019). These approaches rely on parallel data or large monolingual corpora for good quality contextual embeddings. However, for low-resource languages, contextual embeddings from both monolingual and multilingual models are known to be unreliable (Wu and Dredze, 2020).

Later works show the failings of the above approaches in low-resource settings (Adams et al., 2017; Kuriyozov et al., 2020; Chimalamarri et al., 2020; Eder et al., 2021) and propose alternative training strategies such as joint training of static embeddings (Woller et al., 2021; Bafna et al., 2022), and multilingual embeddings from LSTM-based

models (Wada et al., 2019). However, these works either address a higher resource range (>15M tokens), use bilingual lexicons as seeds, or show low scores (≈30 precision@5) for unsupervised BLI. In general, there is a paucity of attention given to setups where there is a severe resource imbalance between the two languages of the BLI pair, despite this being a very typical real-world scenario.

4. Method

Our method is intended for a closely related HRL (source) and LRL (target) pair, written in the same script, and given that we can train or already have a good quality monolingual MLM for the HRL. The main idea is that if we mask an unknown word in the LRL sentence, feed the masked LRL sentence to the HRL MLM, and ask the HRL MLM to propose candidates for the masked LRL word, the HRL MLM should have access to enough contextual cues due to shared vocabulary and syntax to propose meaningful HRL candidates for the masked word. This potentially gives us translation equivalence between the original LRL word and the best scoring proposed HRL candidate. We proceed in an iterative manner, growing the lexicon from equivalents gained from each processed sentence, and using learned equivalents in the lexicon to replace known LRL words with HRL equivalents to process future sentences.

Starting with an empty HRL-LRL bilingual lexicon, we perform the following steps to update our lexicon iteratively, explained in further detail below, and shown in Algorithm 1: (i) we choose an input, consisting of an LRL sentence, and a source LRL word occurring in it, (ii) we replace known words in the input sentence by HRL equivalents using the current state of the lexicon, in order to make the sentence more HRL-like, (iii) the resulting sentence is passed to the HRL MLM to obtain HRL candidate suggestions for the masked LRL word, (iv) we use a reranking heuristic to choose the best HRL candidate, if any, and (v) we update the lexicon if we have found a new equivalent pair.

4.1. Choosing *(sentence, word)* pairs to process

Intuitively, the chance of the HRL MLM giving accurate translation equivalents for the (LRL) word is higher if the LRL sentence is more easily "comprehensible" to the HRL MLM, or if the LRL sentence already has several HRL words in it. Therefore, we aim to first process words in sentences that have a higher concentration of known words, where known words are either shared vocabulary or words that are already in the current state of our lexicon. These words are replaced by their HRL equivalents before the sentence is passed to the HRL MLM.[6] We maintain a priority list of *(sentence, word)* pairs based on the percentage of known words in the sentence and update the list after every batch of sentences based on new learned translations.[7]

4.2. Reranking

The HRL MLM may propose valid candidates for the masked token that are not translation equivalents to the LRL source word; typically, there may be a wide range of reasonable possibilities for any masked word. Therefore, we rerank the returned HRL candidates based on orthographic closeness to the masked LRL word. Our use of orthographic closeness as the basis of our rerankers is motivated by the high percentage of orthographically similar cognates, borrowings, and spelling variants in the vocabulary of these languages with respect to each other (shown by Jha (2019) for Maithili and Hindi). Note that minimum normalized edit distance as a stand-alone approach, i.e. positing the orthographically closest HRL word as a translation equivalent for any LRL word, performs badly for various reasons (Bafna et al., 2022). We compare two rerankers, Basic and Rulebook.

4.2.1. Basic

In the Basic approach, we simply use normalized orthographic similarity (computed using Levenshtein distance) between the candidate and the original masked word. This reranker considers all character substitutions equally costly.

4.2.2. Rulebook

We may see from discovered cognate pairs that certain character transformations are very common (corresponding to regular sound change, or systematic differences in orthographic conventions), and so should be less costly than others. Similarly, different language pairs may have different preferences for cheap or costly character substitutions.

In the Rulebook variant, we use Bafna et al.'s (2022) iterative expectation-maximization (EM) method to learn a custom edit-distance matrix for

[6] We mask whole words and accept single token responses (as the default) from the MLM. In practice, this does not pose a big problem, since the HRL MLM tokenizer has a large vocabulary size (52000): 86% and 81% in the Hindi side of the Bhojpuri and Magahi silver lexicons respectively are preserved as single tokens. We leave it to future work to handle multi-word terms.

[7] Specifically, the priority list is created from the *(sentence, unk_word)* pairs by first sorting them by the number of times each instance has previously been processed, and then by the percentage of other unknown words in the sentence, both in ascending order.

the source and target character sets. This custom edit-distance matrix is used as an orthographic reranker for our approach (lines 6-9 in Algorithm 1).

The idea of this reranker is to iteratively optimize character substitution probabilities from the source to target character set in "known", or hypothesized, cognate pairs, while simultaneously learning new cognate pairs by reranking candidates suggested by the HRL MLM, using the current state of the substitution probabilities.

Setup Let χ_s and χ_t represent the sets of characters on the source (LRL) and target (HRL) sides, respectively. We define a scoring function, $S(c_i, c_j)$ that provides a score for replacing a character $c_i \in \chi_s$ with $c_j \in \chi_t$. Insertions and deletions are considered special cases of replacement, where a null character is introduced or replaced. For a given source set character, S is modelled as a transformation probability distribution over χ_t. Initially, the probabilities in S are assigned to favor self-transformations (typically set to 0.5), and the remaining probability mass is evenly distributed among other characters.

At any given iteration, we can calculate the score for a source-target character substitution, viewed as a conditional probability:

$$S(c_i, c_j) = \frac{C(c_i, c_j)}{T(c_i)} \quad (1)$$

Here, $C(a, b)$ is the number of times we have seen $a \rightarrow b$, and $T(a)$ is the total number of times we have seen a on the source side.

EM Steps for RULEBOOK.

1) Expectation step. Given a list of top k candidates for a given source word s: for each candidate pair (s, t), we find $Ops(s, t)$, which is the *minimal list* of the operations we need to perform to get from s to t. Each member in Ops is of the type (c_i, c_j). Note that we also want to estimate $S(a, a)\ \forall\ a$, and so we also use a "retain" operation, for characters that remain the same. The score for the pair (s, t) is computed as:

$$\zeta(s, t) = -\sum_{(a,b)\in Ops} log(S(a, b)), \quad (2)$$

where the lower the ζ the more probable it is that a pair is equivalent. For a given s, we can then always find the word that is the most probable equivalent as $t_{best} = argmin_{t_i \neq s}(\zeta(s, t_i))$ (line 6 in Algorithm 1). We then add (s, t_{best}) to our learned lexicon (line 8).

2) Maximization step. We update the model parameters based on the newly identified equivalents in the previous step (line 9 in Algorithm 1). This is done by increasing the counts of all observed edit distance operations:

$$C(a, b) := C(a, b) + 1 \quad \forall (a, b) \in Ops(s, t)$$

$$T(a) := T(a) + 1 \quad \forall (a, b) \in Ops(s, t)$$

We disallow updates for $s = t$ (i.e. identical words) in the training phase, to mitigate exploding self-transform probabilities.

4.3. Multiple passes over the input

Once all $(sentence, word)$ pairs have been processed once (or n times), we reprocess them (for an $(n+1)^{th}$ pass) in the hope of gaining more accurate translations, as previously unknown neighbour words may have been learned in the meantime.

4.4. Hyperparameters

We use a minimum normalized orthographic similarity threshold of 0.5 (see line 7 of Algorithm 1). This threshold was heuristically chosen. We set the maximum number of passes to 3, meaning that the algorithm terminates if all unknown words have been processed 3 times. We found in our initial experiments that the algorithm yields very few or no new words in further passes. This also serves as a terminating condition (line 1 in Algorithm 1).

4.5. Examples

We give examples of inputs and outputs of our method in Table 2, illustrating the outputs for BASIC. As we see, the Hindi BERT is fairly good at giving reasonable Hindi candidates for the masked Bhojpuri, although, naturally, these candidates may not be equivalents of the masked word, as shown for the top candidates in rows 2 and 3. Applying reranking based on orthographic similarity solves this problem to a large extent, serving to identify translation equivalents from among given candidates.

We also see an example (row 3) where replacing a Bhojpuri word with its Hindi equivalent in the input sentence helps the Hindi MLM to produce more reasonable Hindi candidates for the masked word.

Algorithm 1: BASIC and RULEBOOK

```
1 while not terminating_condition do
2     sent, word ← chooseLRLExample();
3     sent ← replaceKnownWords(sent, lexicon);
4     sent ← maskWord(sent, word);
5     preds ← HRLBertMaskFill(sent);
6     best ← argmax(orthSim(word, preds));
7     if orthSim(best, word) > threshold then
8         lexicon ← updateLexicon(word, best);
          /* Next step only for RULEBOOK
             */
9         updateOrthSimParams(word, best)
```

#	Source	Listed	Notes	Ideal
1	खाली (only)	केवल (only)	Missing synonym	केवल, सिर्फ़
2	मिलत (meet-1pers)	मिलता (meet-masc.)	Missing fem. inflection	मिलती, मिलता
3	बतवला (share-infinitive)	करने (do-infinitive)	Multi-word equivalence	साझा करने
4	चन्दा (moon)	.	Misc.	चांद

Table 3: Types and examples of faults in the silver lexicon.

Target lang.	#Tokens	Lexicon size	Silver lexicon size
awa	0.17M	10462	-
bho	3.09M	21983	2469
bra	0.33M	10760	-
mag	3.16M	30784	3359
mai	0.16M	12069	-
mar*	551.00M	36929	-
nep*	110.00M	22037	-

Table 4: Monolingual data sizes in tokens, and sizes of our released lexicons (created using our method), and released silver lexicons (from parallel data) for Bhojpuri and Magahi. *High-quality gold bilingual lexicons already exist for these languages.

5. Experimental Settings

Monolingual Data We use monolingual data from the LoResMT shared task (Ojha et al., 2020) for Bhojpuri and Magahi, and the VarDial 2018 shared task data (Zampieri et al., 2018) for Bhojpuri, Awadhi and Braj. For Bhojpuri, we additionally use the BHLTR project (Ojha, 2019). We use the BMM corpus (Mundotiya et al., 2021) and the Wordschatz Leipzig corpus (Goldhahn et al., 2012) for Maithili. For Marathi and Nepali, we use large-scale monolingual corpora made available by IndicCorp (Kakwani et al., 2020) and (Lamsal, 2020) respectively. See Table 4 for monolingual data sizes.

Model We use the MuRIL model and tokenizer (Khanuja et al., 2021) as our HRL MLM for Bhojpuri, Magahi, Awadhi, Maithili and Braj; we use the Hindi BERT and associated tokenizer given by Joshi (2023) for Marathi and Nepali.[8]

Baselines We compare our approaches against semi-supervised VecMap approach with CSLS (Artetxe et al., 2018b,a), using identical words as seeds, with 300-dimensional fastText embeddings (Bojanowski et al., 2017).[9] We also choose CSCBLI (Zhang et al., 2021) as a representative of methods using contextual representations, hypothesizing that the ensemble of static and contextual embeddings may perform better than VecMap. Finally, we report results for a trivial baseline ID, the identity function, representing vocabulary overlap.

Evaluation Data Given the lack of gold lexicons between Hindi and our LRLs, we create silver lexicons instead from parallel data. We use FastAlign with GDFA (Dyer et al., 2013) to extract word alignments from existing gold Bhojpuri–Hindi and Magahi–Hindi parallel data (≈500 sentences per language) (Ojha, 2019).[10] We use the two best candidates per source word in the resulting silver lexicons as valid translations.[11] This yields 2,469 and 3,359 entries for Bhojpuri and Magahi respectively. We report the manually evaluated quality of the silver lexicons in the following paragraph. For Marathi and Nepali, we use existing gold parallel lexicons against Hindi, taken from IndoWordNet (Kakwani et al., 2020), manually aligned to the Hindi WordNet. We obtain lexicons with 35,000 and 22,000 entries for Marathi and Nepali respectively.

Manual Evaluation of Silver Lexicons We perform a manual evaluation of our silver lexicons, in order to judge the credibility of the reported results for our methods for Bhojpuri and Magahi. We manually examine 150 entries in the automatically created Bhojpuri lexicon, and find that 90% of entries are satisfactory, i.e. they list accurate Hindi equivalents of Bhojpuri words. We observe a few general problems with the lexicon, and list representative examples in Table 3:

- Missing common synonyms, e.g. in row 1 of Table 3. This kind of error results in underestimation of precision scores for all approaches.

[8] We need a HRL model that has not seen target data; while MuRIL is a good choice for low-resource dialects because it is multilingual and may benefit from knowledge of other related Indic languages, it cannot be used for Marathi and Nepali because these languages are included in its pretraining data.

[9] Using 100 dimensions gives similar results.

[10] We were unable to find a reasonable quantity of publicly available parallel data for Awadhi, Braj, or Maithili, and so could not perform evaluation for these languages.

[11] This is an empirical choice from eyeballing the resulting silver lexicons; at least one and usually both of the first two translations are valid.

- Problems with correctly equating inflections, missing feminine inflections, e.g. row 2. A natural problem arising from differences in morphological systems of the source and target language is that inflected verbs can be difficult to match cross-lingually. This results in missing equivalents of a given inflected form. For example, while genderless verbs in Bhojpuri should ideally be listed with the corresponding masculine and feminine verbs in Hindi, we observe that they are often missing one gender inflection, usually the feminine one. Similarly, not all possible target inflectional variants of a source inflection are listed for each verb entry.
- Multi-word equivalences lead to errors. For example, in row 3, the single-word Bhojpuri source verb has a noun-light verb complex equivalent in Hindi (consisting of two words, literally meaning "sharing do"), and the silver lexicon lists the light verb ("do") as the target translation. This is also observed in the case of other verb equivalences, where one of the languages using multiple tokens to express an inflection, leading to incorrect matches in the silver lexicon.
- Miscellaneous errors. The lexicon contains some entirely incorrect equivalents (8.76%), due to word alignment errors, e.g. row 4.

Note that we only mark entries as wrong if the listed equivalents are inaccurate, and so faults such as missing synonyms and inflections, which affected 7.33% of the sample we examined, are not represented in the error percentage reported.

6. Results and Discussion

We report precision@2 and accuracy on non-identical predictions (NIA) in Table 5.[12] NIA is calculated by taking all non-identical predictions in the top 2 predictions per word, and reporting the percentage of those predictions that were marked correct by the evaluation lexicons. We report this metric because precision@2 may be inflated by "easy" identical word predictions.

Baselines Table 6 provides examples of the performance of these approaches. VecMap performs well for Marathi: we provide examples where it predicts correct equivalents for rare words (row 8), non-cognates (row 9), as well as frequent words (row 7). However, for Nepali, Bhojpuri, and Magahi, both VecMap and CSCBLI make seemingly random wrong predictions on almost all words (rows 1, 2, and 6), with near-zero performance, probably due to the low quality of static and contextual embeddings for the LRLs. CSCBLI also fails for Marathi, indicating that the Marathi contextual embeddings may still be of poor quality or that the approach may not generalize well to untested language pairs. While the failure of these baselines for Nepali is surprising, it can perhaps be explained by the fact that Nepali has about five times less data than Marathi, and less lexical overlap with Hindi.

Our methods Our Basic and Rulebook approaches outperform ID by more than 20 accuracy points for all languages. Rulebook gains very little, if at all, over Basic, but Rulebook has an edge when it comes to predicting cognates with common sound correspondences (see row 5). We observe that these approaches are reasonably successful for Bhojpuri and Magahi on cognate verbs and common nouns, but fail on syntactic words and postpositions (row 3 for Basic), and may be confused by unrelated words with chance orthographic similarity even for common words (row 5 for Basic). Furthermore, these approaches often predict incorrect inflections of the correct verbal/noun stem (we count these predictions as wrong), as in rows 1 and 4. Although Basic and Rulebook perform with high accuracy for Marathi and Nepali, their NIA is extremely low, indicating that they serve mainly to identify or "sieve" out vocabulary overlap. We see that the candidates proposed by the Hindi MLM are often in fact Marathi/Nepali words, indicating that it has seen some Marathi/Nepali data (due to corpus contamination and/or code-mixing) and is capable of performing mask-filling for Marathi/Nepali.

Manual Evaluation of Generated Lexicons We manually examine errors in the non-identical predictions of Basic, looking at 60 randomly chosen non-identical Bhojpuri predictions.[13] We find that 31.7% of predictions are correctly inflected equivalents, as opposed to 18.1% given by the NIA quantitative evaluation. The underestimation is caused by missing synonyms in the silver lexicon. Furthermore, 25% are incorrectly inflected cognates of the source word, and the rest are unrelated words.

How useful is reranking by orthographic distance? We also ran the Basic approach without reranking with orthographic distance, i.e. we simply pick the top candidate suggested by the HRL mask-filling model as an equivalent. This approach is clearly worse than the standard Basic approach

[12] We also report P@{1,3,5} in Appexdix A; these results show similar trends.

[13] One of the authors is a native speaker of Hindi but not Bhojpuri. We segregate translation equivalents into error categories using (self-made) inflection tables inferred from the silver lexicons, as well as cognate knowledge from Hindi native speaker knowledge.

		bho		mag		mar		nep	
	Method	P@2	NIA	P@2	NIA	P@2	NIA	P@2	NIA
Baselines	ID	37.3	0.0	39.9	0.0	27.5	0.0	21.2	0
	VecMap+CSLS	0.0	0.0	1.2	0.6	42.4	**26.7**	0.0	0.0
	CSCBLI	0.0	0.0	2.0	0.5	0.0	0.0	0.0	0.0
Ours	Basic	61.0	**18.1**	65.2	**18.8**	**80.9**	2.8	**87.6**	**8.2**
	Rulebook	**61.5**	15.1	**65.4**	17.4	80.6	1.72	**87.6**	6.0

Table 5: Performance of the methods, given by Precision@2 (P@2) and accuracy of non-identical predictions (NIA).

#	Lang	Word	Correct	Basic	Rulebook	VecMap	CSCBLI
1	bho	देखत (sees)	देखता	देख†	देख†	अटपटे (weird)	मंत्रमुग्ध (spellbound)
2		मिलत (meets)	मिलते	मिलते	मिल†	गा (sing)	गा (sing)
3		इहाँ (here)	यहाँ	इतिहास (history)	यहाँ	लहरी (wavy)	नजारा (view)
4	mag	डालऽ (puts)	डालती	डाले†	डाल†	तुने*	बहुतों (many)
5		सबाल (question)	सवाल	बोल (speak)	सवाल	विधायिका*	विधायिका*
6		चोरा (steal)	चुरा	चोरी†(theft)	चोर†(thief)	दिहाड़े (day)	दिहाड़ी (day)
7	mar	थंडी (cold)	ठंड	थंडी	थंडी	ठंड	ज्योति (light)
8		किमान (at least)	न्यूनतम	किमान	किमान	न्यूनतम	swift
9		अनादर (disrespect)	अपमान	अनादर	अनादर	अपमान	चामुंडेश्वरी (place name)

Table 6: Predictions made by different approaches. Meanings are provided for the first occurrence of the word. * indicates a non-word and † a prediction in the wrong inflectional/derivational form of the target.

(with reranking), performing at only 3.03% NIA for Bhojpuri and 4.04% NIA for Magahi (approximately -15 and -14 percentage points compared to BASIC for Bhojpuri and Magahi respectively, as shown in Table 5). However, this approach can still identify and capture identical vocabulary.

Variants We experimented with minor variants of the RULEBOOK update mechanisms to see if they result in boosts to performance. We tried disallowing updates for the null character, since we found that a large probability mass iteratively accumulates in the null character (or for deletion). We also incorporated a change in the original algorithm, whereby we made updates to the custom edit distance matrix based on the optimal list of substitutions as per the current state of the edit distance matrix, rather than choosing a minimal length path at random (with each substitution counted as length 1) from the source to the target word when several exist. However, these variants result in very minor improvements or even slight degradations to performance, and we do not report these results.

7. Details of released lexicons

We make our bilingual lexicons publicly available under a CC BY-NC 4.0 license for Bhojpuri, Magahi, Awadhi, Braj, and Maithili, and also release our created silver evaluation lexicons for Bhojpuri and Magahi under the same license. These are the first publicly available bilingual lexicons all these languages except Bhojpuri, to the best of our knowledge. The sizes of the released lexicons for each target language are provided in Table 4. Note that while we also release our generated lexicons for Marathi and Nepali, large high quality gold bilingual lexicons already exist for these languages (see Section 5) and should be used instead of ours; we are mainly interested in creating resources for the low-resource languages.

8. Conclusion

We introduce a novel method for unsupervised BLI between a related LRL and HRL, which only requires a good quality MLM for the HRL. This addresses an important gap in the existing literature, which often relies on good quality embeddings for both languages. Our method shows superior performance on two low-resource languages from the Indic continuum, against near-zero performances of existing state-of-the-art methods. We perform control experiments for two more distantly related Indic languages, and release resulting bilingual lexicons for five truly low-resource Indic languages.

Limitations

The applicability of our method is restricted to low-resource languages that are related to a high-resource language. As the BASIC and RULEBOOK

method are directly dependent on orthographic distance between translation pairs, they are only useful for identifying cognate equivalents, borrowings, or alternate spellings in the source and target language. We also clarify that our method is not intended for mid-to-high resourced language pairs (such as Marathi–Hindi), where canonical state-of-the-art methods such as VecMap work more robustly, specifically on non-identical word equivalents. Our method therefore has a specific (although important) target scenario, i.e. it is a simple method to build bilingual lexicons for severely under-resourced languages leveraging the resources of a closely related high-resource language, given that state-of-the-art methods fail in these settings. Note that we also only deal in the entirely unsupervised scenario in keeping with typical conditions for our target languages (see Section 2), and leave it to future work to improve these methods with a little supervision from bilingual lexicons, possibly obtained from parallel data.

Another limitation of our work is that we were not able to provide true native speaker evaluation for the resulting target language lexicons, instead providing evaluation by the first author (Hindi native speaker) relying on knowledge of shared cognates, the morphology of the target language, and inflection tables. We provide examples in Table 6 and Table 2, and release the automatically created as well as silver lexicons. Finally, our method is only capable of providing single token (HRL) matches to the masked (LRL) whole word. As discussed in Section 4, this problem does not affect the large majority of cases. We leave it to future work to extend our idea to handle multi-token words and multi-word expressions using, for example, span-filling language models (Donahue et al., 2020).

Ethics Statement

Our work is driven by the aim to boost NLP for severely under-resourced languages of the Indic language belt, as well as contribute a method that may be relevant to other language families with a similar linguistic and resource setup. Our method relies on the predictions of language models for the high-resource language and is therefore fallible to general ethical issues with such models, including caste, religion, and gender biases shown to be exhibited by such models (Malik et al., 2022).

Acknowledgements

This work was partly funded by the last two authors' chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. First and second authors are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) —Project–ID 232722074— SFB 1102. Second and third authors are supported by the EU project LT-Bridge (GA952194).

9. Bibliographical References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning Principled Bilingual Mappings of Word Embeddings While Preserving Monolingual Invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning Bilingual Word Embeddings with (Almost) No Bilingual Data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual Lexicon Induction through Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Niyati Bafna, Josef van Genabith, Cristina España-Bonet, and Zdeněk Žabokrtský. 2022. Combining Noisy Semantic Signals with Orthographic Cues: Cognate Induction for the Indic Dialect Continuum. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 110–131, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological Segmentation to Improve Crosslingual Word Embeddings for Low Resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(5):69:1–69:15.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. *arXiv preprint: arXiv:1710.04087*.

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

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic Transfer Using a Bilingual Lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11, Jeju Island, Korea. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Tobias Eder, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based Bilingual Word Embeddings for Low-Resource Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 227–232, Online. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2020. Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Sanjay Kumar Jha. 2019. Exploring the Degree of Similarities between Hindi and Maithili Words from Glottochronological Perspective. *International Journal of Innovations in TESOL and Applied Linguistics*, 5.

Raviraj Joshi. 2023. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint: arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages.

Elmurod Kuriyozov, Yerai Doval, and Carlos Gómez-Rodríguez. 2020. Cross-Lingual Word Embeddings for Turkic Languages. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 4054–4062, Marseille, France. European Language Resources Association.

Rabindra Lamsal. 2020. A Large Scale Nepali Text Corpus.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially Aware Bias Measurements for Hindi Language Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint: arXiv:1301.3781*.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2021. Linguistic Resources for Bhojpuri, Magahi, and Maithili: Statistics about Them, Their Similarity Estimates, and Baselines for Three Applications. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20:1–37.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. HABLex: Human Annotated Bilingual Lexicons for Experiments in Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1382–1387, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124, Florence, Italy. Association for Computational Linguistics.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do Not Neglect Related Languages: The Case of Low-Resource Occitan Cross-Lingual Word Embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive Refinement of Cross-Lingual Word Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996, Online. Association for Computational Linguistics.

Jinpeng Zhang, Baijun Ji, Nini Xiao, Xiangyu Duan, Min Zhang, Yangbin Shi, and Weihua Luo. 2021. Combining Static Word Embeddings and Contextual Representations for Bilingual Lexicon Induction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2943–2955, Online. Association for Computational Linguistics.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross Language Dependency Parsing using a Bilingual Lexicon. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 55–63, Suntec, Singapore. Association for Computational Linguistics.

10. Language Resource References

Goldhahn, Dirk and Eckart, Thomas and Quasthoff, Uwe. 2012. *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages*. European Language Resources Association (ELRA). PID https://aclanthology.org/L12-1154/.

Kakwani, Divyanshu and Kunchukuttan, Anoop and Golla, Satish and Gokul, NC and Bhattacharyya, Avik and Khapra, Mitesh M and Kumar, Pratyush. 2020. *inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. PID https://indicnlp.ai4bharat.org/home/.

Lamsal, Rabindra. 2020. *A Large Scale Nepali Text Corpus*. IEEEdataport. PID https://doi.org/10.21227/jxrd-d245.

Mundotiya, Rajesh Kumar and Singh, Manish Kumar and Kapur, Rahul and Mishra, Swasti and Singh, Anil Kumar. 2021. *Linguistic Resources for Bhojpuri, Magahi, and Maithili: Statistics about Them, Their Similarity Estimates, and Baselines for Three Applications*. ACM Transactions on Asian and Low-Resource Language Information Processing. PID https://github.com/singhakr/Bhojpuri-Magahi-and-Maithili-Linguistic-Resources.

Ojha, Atul Kr. 2019. *English-Bhojpuri SMT System: Insights from the Karaka Model*. arXiv. PID http://arxiv.org/abs/1905.02239.

Ojha, Atul Kr. and Malykh, Valentin and Karakanta, Alina and Liu, Chao-Hong. 2020. *Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low-Resource languages*. Association for Computational Linguistics. PID https://aclanthology.org/2020.loresmt-1.4.

Zampieri, Marcos and Nakov, Preslav and Ljubešić, Nikola and Tiedemann, Jörg and Malmasi, Shervin and Ali, Ahmed. 2018. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics. PID https://aclanthology.org/W18-3900.

	bho			mag		
	P@1	P@3	P@5	P@1	P@3	P@5
VecMap+CSLS	0	0	0	1.2	1.2	1.2
Basic	58.1	61	61	62.5	65.1	65.1
Rulebook	**59.1**	**61.6**	**61.6**	**63**	**65.4**	**65.4**

Table 7: P@{1,3,5} for `bho` and `mag`.

	bho			mag		
	NIA@1	NIA@3	NIA@5	NIA@1	NIA@3	NIA@5
VecMap+CSLS	0	0	0	0.6	0.6	0.6
Basic	**23.7**	**17.1**	**17.1**	**27.2**	**18.4**	**18.2**
Rulebook	20.2	14.4	14.5	23.1	17	16.9

Table 8: NIA@{1,3,5} for `bho` and `mag`.

A. Additional Results

We report P@1,3,5 in Table 7 and NIA@1,3,5 in Table 8. We see that both BASIC and RULEBOOK approaches do not benefit from considering more than 3 best answers. In general, we see the same relative trend as in Table 5.