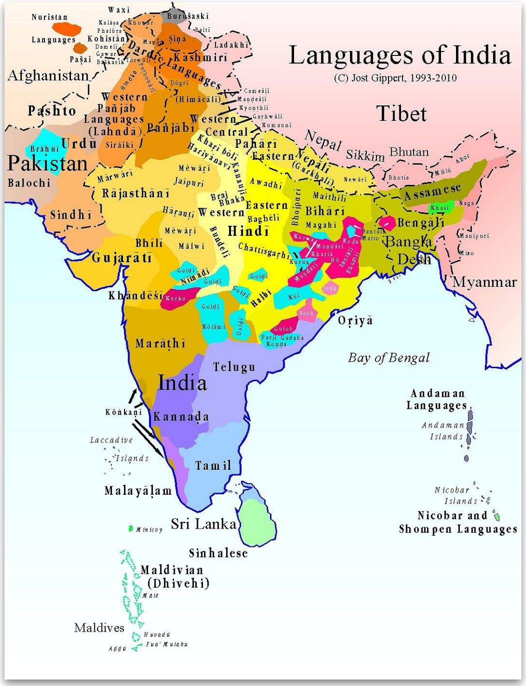
When Your Cousin has the Right Connections: Unsupervised Bilingual Lexicon Induction for Related Data-Imbalanced Languages

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

Unsupervised BLI i.e. what to do when you have

- 576 self-reported mother tongues, grouped into 121 languages[1]
- 15-22 mid-resource languages with official status
- High demand
- Limited funding and interest in data collection
  - Some monolingual data but not enough for good static/contextual embeddings
- Arabic continuum, Turkic continuum

[1] https://censusindia.gov.in/



Lexical relationships between CRLs

- We work with the Indic dialect continuum
- 40+ closely related languages
- High number of shared cognates with Hindi

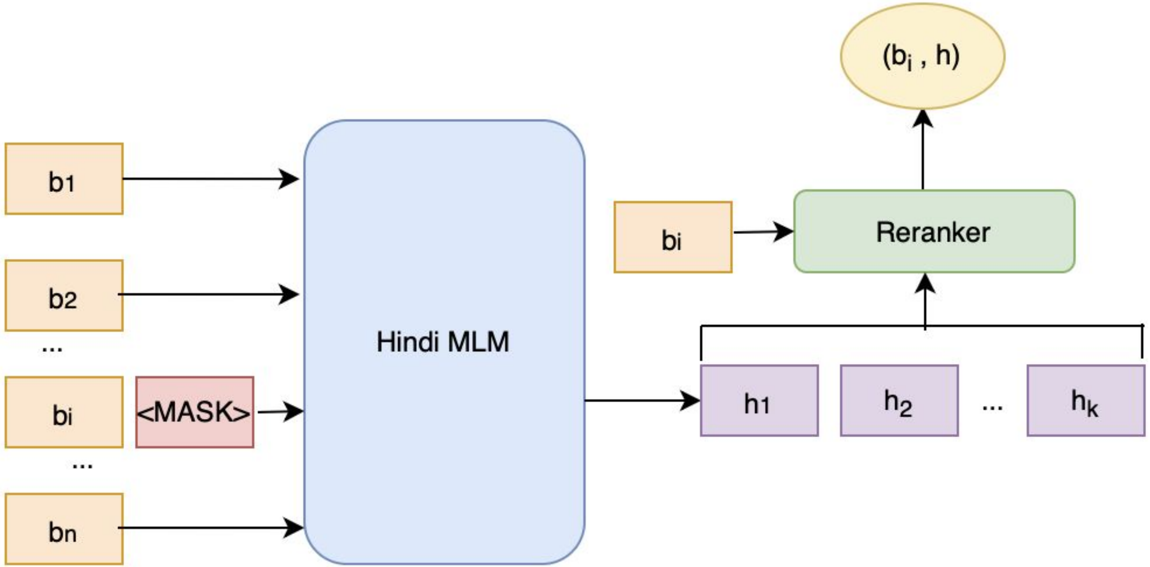
Meaning	boy (nom)	sister (nom)	your (hon., fem. sing. obj)	told (completive)	(you) are going
Hindi	ləɖkaː	bəhən	aːpkiː	bət̪ːaːjaː/ kəːh lijaː	dʑaː rəheː hoː
Awadi	ləɖkaː	bəhin	aːpən	bət̪ːaːvət̪	dʑaːt̪ əhaːi
Bhojpuri	ləikaː	bəhin	aːpən	kəhəl	dʑaːt̪ baː
Magahi	ləiːkaː	bəhin	əpən	kəhəlieː	dʑaː həi
Maithili	ləɖkaː	bəhin	əhaːnk	kəhəlhu^{n}	dʑaː rəhəl ətʃ^{h} 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).

Method

Main Idea: Using Hindi MLMs to extract cognates

- Hindi MLM can do masked word prediction on LRL text
- Produced Hindi candidates may contain translation equivalents of masked LRL word



Reranking

- MLM produces several candidates with probabilities
  - May be semantically correct options that are not cognates/equivalents of source

- We rerank these using orthographic similarity to the source word
  - Motivated by high percentage of cognates, spelling variants, and borrowings across these languages
  - Assumes shared script

Basic

- Reranks with normalized Levenshtein distance between source and target
- Treats all character substitutions equally
- Treats all language pairs in the same way

Rulebook: Learning Custom Levenshtein Matrices

- Learn a custom Levenshtein matrix using an EM approach[1]

- Iterative approach:
  - (E-Step) Find new cognates based on existing char sub scores
  - (M-Step) Estimate scores for char substitutions based on existing cognate list

- Initialization:
  - High prob to self transform (retention)
  - Distribute prob mass to other chars

- Estimation:
  - Score for single char pair: Frequentist prob.
  - Score for cognate pair: Minimal operations list (product)
  - Updating scores: increment counts

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

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

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

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

[1]Taken from (Bafna et al., 2022)

Priority Processing

- MLM will do better if it already knows most of the words in the sentence (i.e. they are in Hindi)
  - Initially, we will rely on shared vocabulary

- Once we obtain a translation pair *(b, h)*, we replace all instances of *b* in input LRL text with *h*

- We process input LRL (sentence, word) pairs in priority determined by percentage of shared/known words in the sentence
  - May be shared vocabulary i.e. present in Hindi vocabulary
  - Or LRL word for which we know the Hindi translation

Input and Output Examples for Bhojpuri		
1	Input Mask Correct Preds	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.’ भरल ‘filled’ भरी परिपूर्ण, <u>भरी</u>, युक्त, भरपूर, सम्पन्न replete, filled, containing, filled-up, prosperous
2	Input Mask Correct Preds	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.’ कइलन ‘did’ की, करी करे, करी, की, किया, <u>*करेल</u> do-hypothetical, did-fem, did-fem, did-masc, -
3	Input Mask Correct Preds New input Preds	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.’ लोगन ‘people’ लोग, लोगों बात, काम, <u>लड़की</u>, दिन, औरत thing, work, girl, day, woman हमनी के **उन** [MASK] पर बहुते गर्व बा । सब, लोग, <u>लोगों</u>, दिन, सभी all of (them), people, people, day, all of (them)

Experimental Setup

Data

- Monolingual data from LoResMT (Ojha et al., 2020) : Bhojpuri, Magahi
- VarDial 2018 shared task data (Zampieri et al., 2018): Bhojpuri, Awadhi and Braj
- BHLTR project (Ojha, 2019): Bhojpuri
- BMM corpus (Mundotiya et al., 2021) : Maithili
- Wordschatz Leipzig corpus (Goldhahn et al., 2012): Maithili
- IndicCorp (Kakwani et al., 2020): Marathi
- (Lamsal, 2020): Nepali

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.

Models

- Need an HRL model that has *not* seen LRL data (since we want Hindi equivalents)

- MuRIL model and tokenizer (Khanuja et al., 2021) for Bhojpuri, Magahi, Awadhi, Maithili and Braj
  - LRLs may benefit from other language data in pretraining corpus

- Hindi BERT and associated tokenizer (Joshi et al., 2023) for Marathi and Nepali

Baselines

- Semi-supervised VecMap with CSLS (Artetxe et al., 2018)
  - Identical words as seeds
  - 100, 300 dimensional fastText embeddings

- CSCBLI (Zhang et al., 2021) - representative of methods using static and contextual embeddings
  - Uses spring network to align non-isomorphic contextual embeddings
  - Interpolates with static embeddings
  - Comparable/superior results to other methods using contextual embeddings

Evaluation Data

Evaluation lexicons

- For Marathi, Nepali, we use gold lexicons from IndoWordNet (Kakwani et al., 2020)
  - Manually aligned to Hindi WordNet

- For Bhojpuri and Magahi, we create silver lexicons
  - ~500 parallel sentences with Hindi (Ojha, 2019)
  - FastAlign with GDFA
  - 2469 Bhojpuri entries, 3359 Magahi entries

Evaluation of Silver Lexicons

- Manual evaluation of 150 entries : 90% entries are accurate

- Problems in the lexicon:

(1) Missing synonyms
(2) Missing female inflections, wrong inflections
(3) Errors with multiword equivalences
(4) Misc.

#	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.

Results and Discussion

Automatic Evaluation

		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).

Automatic Evaluation

- We report P@2 (also P@1,3,5 in paper)
- NIA: accuracy on non-identical pairs (since identical pairs are easy)

- VecMap and CSCBLI:
- Works best for Marathi (on frequent words, rare words, non-cognates)
- Seemingly random predictions on other languages

- Basic, Rulebook give ~20 pt gains for Bhojpuri, Magahi
- Successful on cognate verbs and nouns, fail on functional words
- Can be confused by chance orthographic similarity
- Predict incorrect inflections

Examples

#	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.

Manual Evaluation

- Manually examine 60 non-identical predictions from Bhojpuri test set
- 31.7% P@2 (automatic evaluation underestimates due to missing synonyms)
- 25% incorrect inflections
- Rest unrelated words

Released Lexicons

- Generated lexicons for Bhojpuri, Magahi, Maithili, Awadhi, Braj made available

- Silver evaluation lexicons for Bhojpuri and Magahi made available

- https://github.com/niyatibafna/BLI-for-Indic-languages.

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.

Conclusion

Takeaways

- In extremely low-resource scenarios, embeddings-based approaches break completely
- And we need more robust and less data hungry approaches

- We can make LMs trained on a closely related cousin read LRL text and give us potential cognate equivalents for masked words

- We use further reranking tricks to filter candidates
- Relying on orthographic similarity - directly, custom Levenshtein matrices

- These approaches outperform embeddings-based approaches by a wide margin

- Plenty of work to be done to improve absolute performance in these scenarios

- Check out our released lexicons!

Thank you! :-)

Other things

More results

- No reranking (pick top MLM candidate)
- -15, -14 pts compared to Basic for Bhojpuri, Magahi

Notes

- We do several iterations over the corpus since we have learnt new context words in the meantime
  - May get better translations for previously processed LRL word
  - We find that no new words are updated after ~3 iterations

- We use empirically determined thresholds for rerankers
  - For a given input, we may find that all MLM candidates are bad
  - In this case, we add nothing to the lexicon

- For Rulebook, initially source-target distributions are set to favour identity (0.5 probability mass)