A Generalized Method for Automated Multilingual Loanword Detection

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy

Colorado State University — Situated Grounding and Natural Language Lab

abhijnan.nath@colostate.edu • sina@cs.utah.edu • {ibrahim.khebour,sheikh.mannan,nkrishna}colostate.edu

Introduction

- **Loanwords:** words incorporated from one language to another without translation
- If two words sound similar and have similar meanings, this is (usually) too coincidental to have occurred by chance
- We present a method to automatically detect loanwords between arbitrary language pairs
- Account for phonetic, semantic, orthographic, and articulatory features
- Evaluate on 12 language pairs, 4 unseen language pairs
- Our method achieves or exceeds SOTA and human performance
- Findings suggest features of loanwords allow generalization

Data

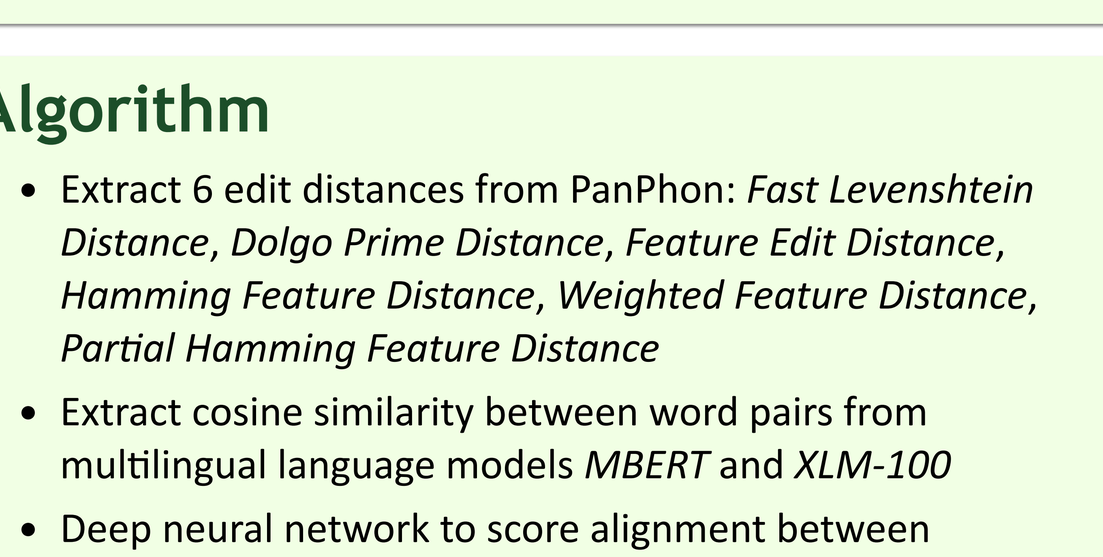
- Wiktionary LoanWord (WikLoW) dataset: 16 language pairs gathered from Wiktionary, with extensible method
- Positive loans augmented with:
  - **Synonyms** (similar meaning, different pronunciation)
  - **Hard Negatives** (different meaning, similar pronunciation)
  - **Randoms** (different meaning, different pronunciation)
- Converted to IPA using Epitran and articulatory features using PanPhon

Borrower	Donor	# loans
English	French	5074
English	German	2942
Indonesian	Dutch	2665
Polish	French	2055
Romanian	French	2000†
Kazakh*	Russian	1809
Persian	Arabic	1526
Romanian	Hungarian	1460
German	French	1365
Hindi*	Persian	1249
Finnish	Swedish*	1242
Azerbaijani*	Arabic	1116
Mandarin	English	960
Hungarian	German	532
German	Italian	249
Catalan*	Arabic	94

Loanword counts per language pair

Algorithm

- Extract 6 edit distances from PanPhon: *Fast Levenshtein Distance*, *Dolgo Prime Distance*, *Feature Edit Distance*, *Hamming Feature Distance*, *Weighted Feature Distance*, *Partial Hamming Feature Distance*
- Extract cosine similarity between word pairs from multilingual language models *MBERT* and *XLM-100*
- Deep neural network to score alignment between articulatory features
- **Binary classification:** Logistic Regressor, Neural Network, Support Vector Machine, Random Forest



Loanword detection architecture

Results

	LR	NN	SVM	RF
F1 (+)	85	86	84	85

Avg. positive F1 (%) of 4 classifiers

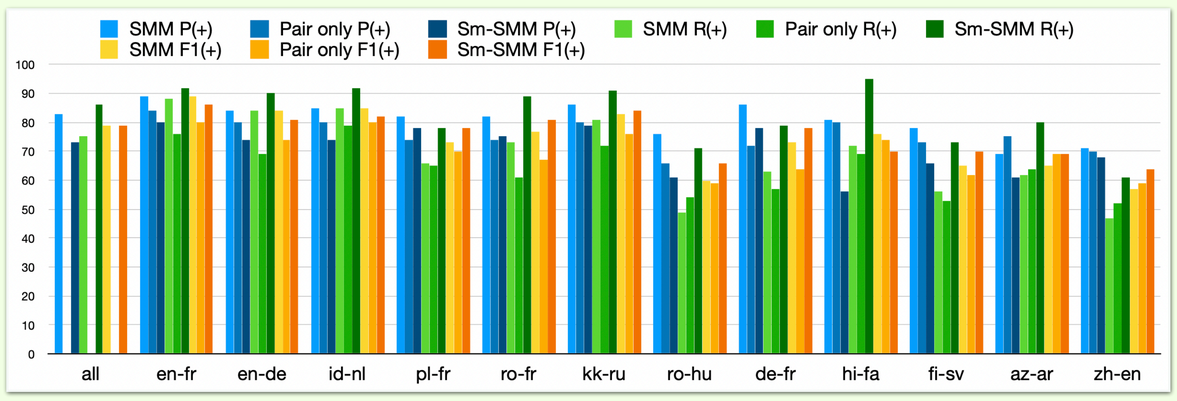
		all	*en-fr*	*en-de*	*id-nl*	*pl-fr*	*ro-fr*	*kk-ru*	*ro-hu*	*de-fr*	*hi-fa*	*fi-sv*	*az-ar*	*zh-en*
P (+)		92	96	90	96	90	94	93	88	94	94	85	85	81
		98	97	98	99	97	96	98	99	98	97	98	98	98
		83	89	84	85	82	82	86	76	86	81	78	69	71
R (+)		81	91	87	90	73	82	88	61	75	86	68	71	51
		98	99	99	99	97	99	100	93	99	99	98	98	93
		75	88	84	85	66	73	81	49	63	72	56	62	47
F1 (+)		86	93	89	93	81	88	91	72	83	90	75	70	62
		98	98	98	99	97	98	99	96	98	98	98	98	95
		79	89	84	85	73	77	83	60	73	76	65	65	57

SMM neural network results (%) (1st row: realistic distribution of loanwords; 2nd row: balanced between loans/non-loans; 3rd row: all available data)

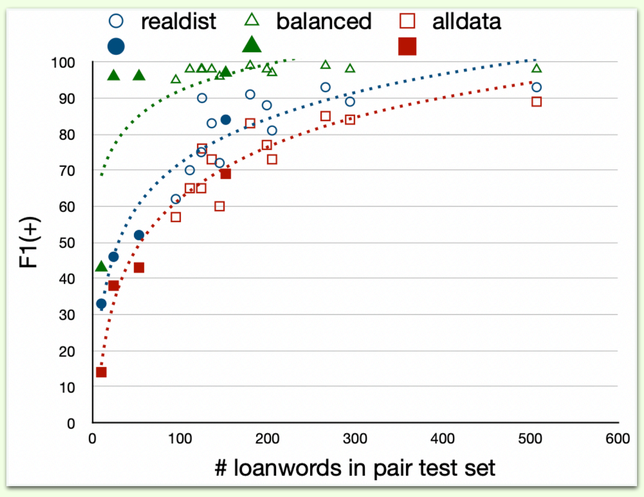
		fa-ar	*hu-de*	*de-it*	*ca-ar*
P (+)		95	95	73	100
		97	100	100	75
		75	73	54	25
R (+)		75	36	33	20
		97	93	92	30
		64	30	29	10
F1 (+)		84	52	46	33
		97	96	96	43
		69	43	38	14

SMM performance on 4 unseen pairs (%)

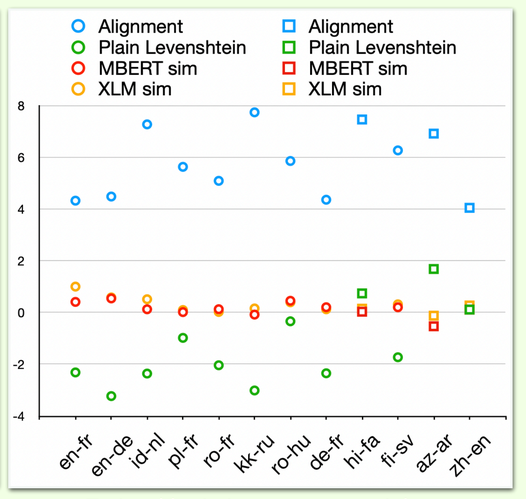
- Four experiments: *Single Multilingual Model (SMM)*, *pair-specific models*, *pruned training set (Small-SMM)*, *unseen language pairs*



Neural network results comparing SMM, pair-specific models, and Small-SMM



F1 score vs. number of loans per pair



Trained logistic regressor weights

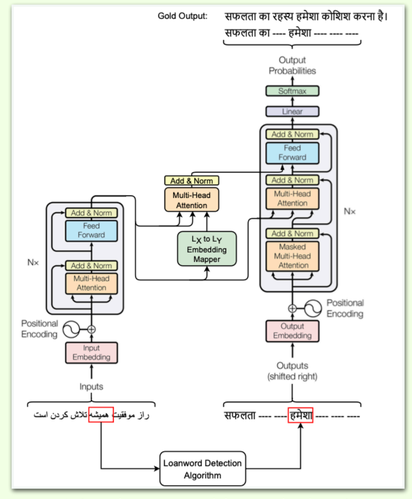
- Our method generalizes to unseen language pairs
- Articulatory alignment most useful feature
- Human comparison: fluent speakers selected loans from same test set
- Model can beat human performance too!

Pair	N	Human μ R(+)	SMM R(+)	κ	# loans (homonyms)
en-fr	7	29	**88**	.059	508 (8)
hi-fa	6	60	**72**	.113	125 (4)
zh-en	6	**85**[9]	47	.034	95 (1)

Human recall vs. SMM recall

Conclusion and Future Work

- We present an extensible method and novel baseline in loanword detection for arbitrary language pairs
- Automated loanword detection enables many downstream tasks
- Loanword knowledge is useful in, e.g., coreference resolution, NER, MT
- Parallel loanword corpora afford learning cross-lingual embedding mappings



Using loanword knowledge in machine translation

Resources

- Codebase: https://github.com/csu-signal/loan-word-detection