Faculty of Humanities
School of Linguistics

Linguistic Convergence Laboratory

Moscow
2022

Automatic Detection of Borrowings in Low-Resource Languages of the Caucasus: Andic Branch

Authors: Konstantin Zaitsev, Anzhelika Minchenko

Task & motivation

Loanwords occur in all languages, but their detection can be problematic in low-resource languages.

Our main goal is to detect borrowings automatically without using a bilingual dictionary. Automation can facilitate future field research on target languages.

Reasons for the study

1. the lack of their writing system or stable spelling;
2. lack of qualified linguists and translators for the given language;
3. limited distribution on the Internet;
4. lack of electronic resources for language and speech processing, including monolingual corpora, bilingual electronic dictionaries, spelling and phonetic transcriptions of speech, pronunciation dictionaries, and more.

Dataset

Statistics of Andic languages

Glottocode	Language	Number of Words
akhv1239	Akhvakh	14007
andi1255	Andi	6144
bagv1239	Bagvalal	12706
botl1242	Botlikh	21483
cham1309	Chamalal	9721
ghod1238	Godoberi	7423
kara1474	Karata	6650
tind1238	Tindi	12419

Dictionary description for the Akhvakh language

lemma	ipa	glottocode	bor	borrowing_source_language	meaning_ru
аба'далIи	a-b-'a-d-a-t-ɬː-i	akhv1239	1	arab	Eternal
а/б/а'жве	a-b-'a-ʒʷ-e	akhv1239	0	NaN	everlasting
а/б/ажу'рулъIа	a-b-a-ʒ-'u-r-u-t-ɬ-a	akhv1239	0	NaN	communicate

Baseline

We used:

1. Logistic regression;
2. Tfldf-vectorizer.

Language	Precision	Recall	F1
Ahvakh	0.90	0.57	0.60
Andi	0.80	0.56	0.58
Bagvalal	0.81	0.60	0.63
Botlikh	0.88	0.74	0.78
Chamalal	0.97	0.51	0.50
Godoberi	0.89	0.61	0.65
Karata	0.96	0.51	0.49
Tindi	0.97	0.53	0.54

Improving baseline

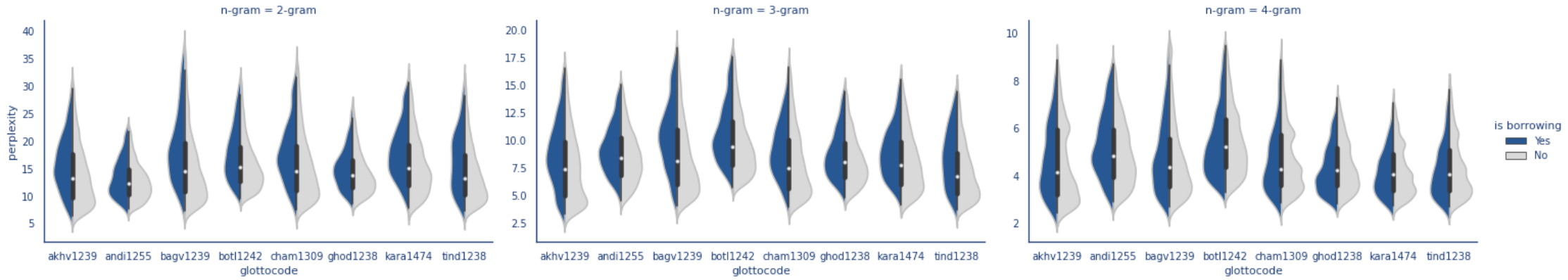
Changes:

1. Switched TfIdf-vectorizer to CountVectorizer;
2. Selected hyperparameters for vecrorizer (min_df = 0.001, max_df = 0.1);
3. Selected the most important features/n-grams.

Language	Precision	Recall	F1
Ahvakh	0.79	0.72	0.74
Andi	0.75	0.69	0.71
Bagvalal	0.80	0.71	0.74
Botlikh	0.86	0.83	0.85
Chamalal	0.80	0.65	0.70
Godoberi	0.82	0.77	0.79
Karata	0.76	0.65	0.69
Tindi	0.73	0.65	0.68

Language model approach

We implemented a language model using Markov chains on n-grams. For the model, we developed the perplexity metric. The metric shows how the word corresponds to a language.



Feature extraction algorithm

To extract features from a word, we used an algorithm. It works as follows:

1. Each input word is divided into n-gram;
2. N-grams is checked in the language model:
    1. If n-gram is not in the language model, then we add a positive coefficient;
    2. Otherwise, we add a negative coefficient;
3. We compute sum of word coefficients and divide them the word length.

Combining models

We combined the improved baseline and the feature extraction algorithm.

Language	Precision	Recall	F1
Ahvakh	0.75	0.83	0.78
Andi	0.72	0.76	0.74
Bagvalal	0.78	0.82	0.80
Botlikh	0.80	0.88	0.83
Chamalal	0.76	0.80	0.78
Godoberi	0.78	0.86	0.81
Karata	0.70	0.73	0.71
Tindi	0.70	0.77	0.73

Results

Model quality comparisons

Model	Akhvakh	Andi	Bagvalal	Botlikh	Chamalal	Godoberi	Karata	Tindi
Baseline	0.60	0.59	0.63	0.78	0.50	0.65	0.50	0.54
BF	0.73	0.69	0.74	0.84	0.68	0.78	0.68	0.66
BFLMipa	0.78	0.74	0.80	0.83	0.78	0.81	0.71	0.73
BFLMlem	**0.82**	**0.77**	**0.84**	**0.86**	**0.79**	**0.84**	**0.75**	**0.75**

Our models' results compare to other research

Model	Precision	Recall	F1
our BFLMipa	0.75	0.81	0.77
our BFLMlem	0.78	**0.83**	0.80
Neural Network for Uyghur	0.82	0.79	0.80
BiLSTM-CRF for Spanish	**0.91**	0.79	**0.84**

Discussion and future research

1. The dictionary does not fully reflect the quality of the. For this reason, the model must be tested on work with texts;
2. The model works in terms of binary classification. In the future, we may refine the model adding definition of the source-language of the borrowing;
3. We expect our findings might be used in other models solving a borrowing detection task. Also detected borrowings might be helpful for field linguists to understand deeply these languages.

Contacts:

konstantzts@gmail.com anzhelika.min@gmail.com

@adugeen @howtouns_s