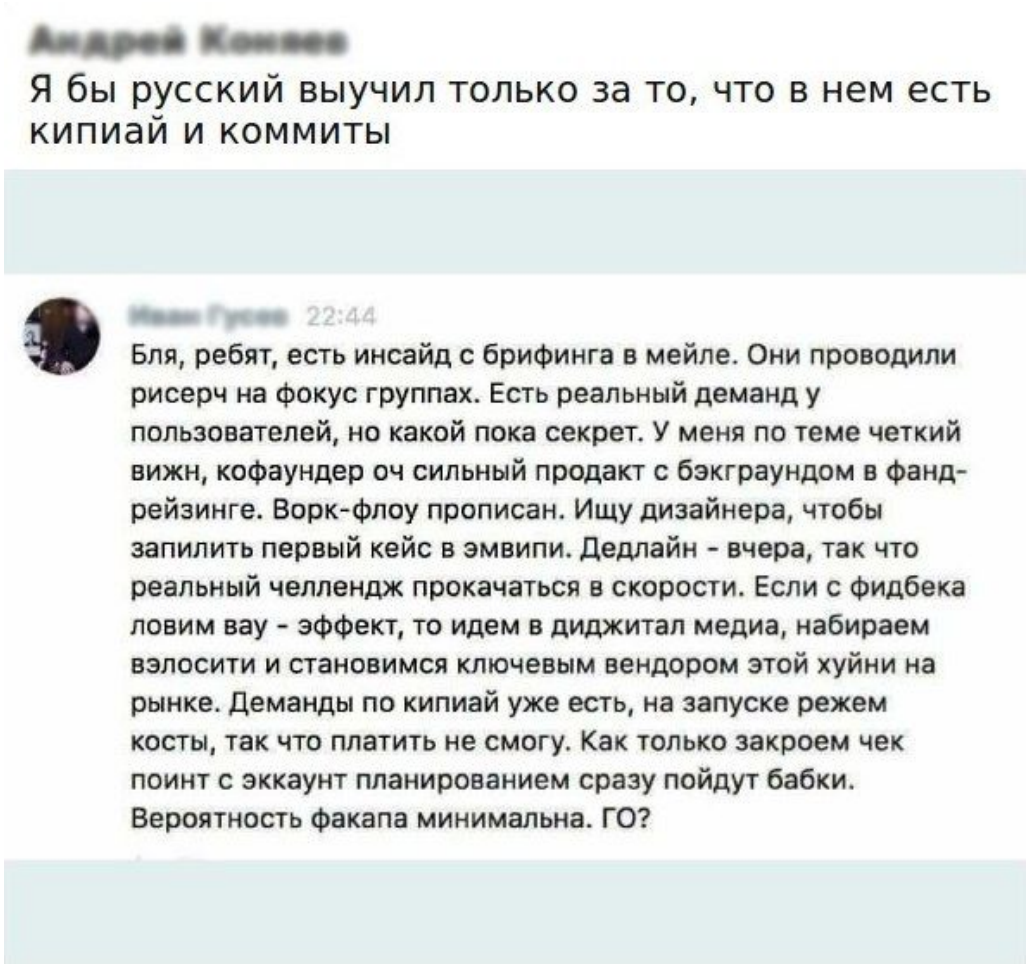
A General Method Applicable to the Search for Anglicisms in Russian Social Network Texts

Alena Fenogenova, Ilia Karpov, Viktor Kazorin

National Research University Higher School of Economics,
Scientific Research Institute KVANT
AINL 2016

Problem

The phenomenon of anglicisms in Russian language



Task

Growing number of loanwords in Russian:

- problem for NLP tasks like spell-checking, taggers, sentiment detection, etc.
- interest for theoretical researches, studying language contact processes

We propose the method for detecting English loanwords (anglicisms) in Russian Social Network texts

- unsupervised
- fully automated
- can be applied to any domain-specific area

Methodology

Idea: simultaneous scripting, phonetics and semantics similarity of the original Latin word and its Cyrillic analogue

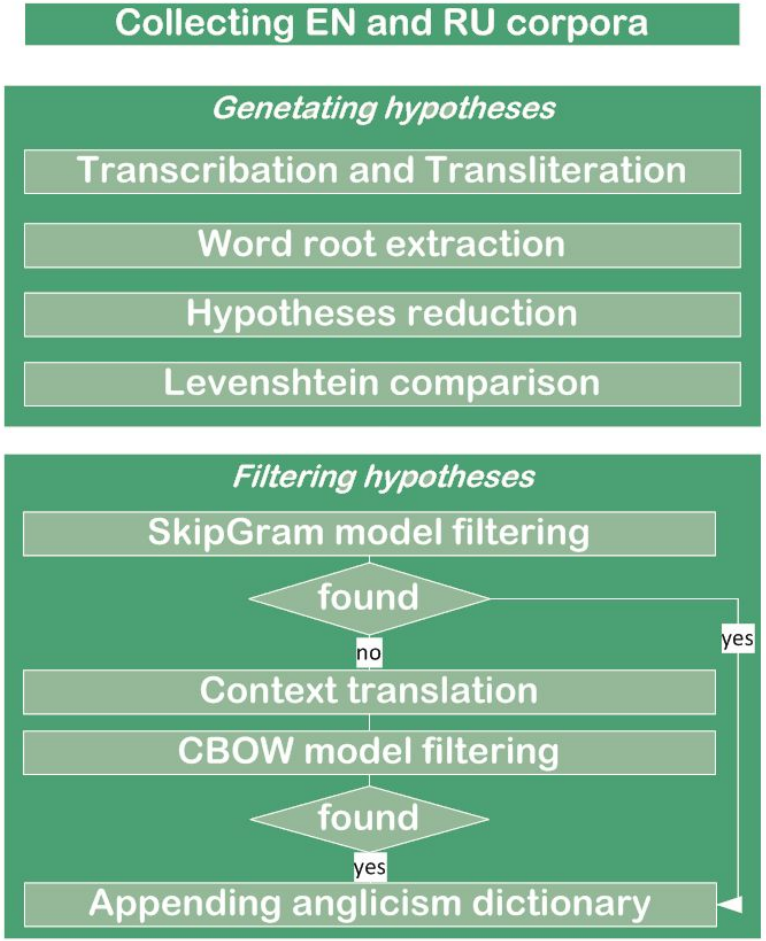
Hypotheses generation:

> Set of transliteration, phonetic transcribing and morphological analysis methods to find possible hypotheses

Hypotheses filtering:

> Distributional semantic models for filtering hypotheses

General Architecture



Corpus collection

LiveJournal blog platform:

- English and Russian
- 20.000.000 texts and comments
- from 100,000 Russian and English top bloggers
- large variety of themes
- users of different age
- less plagiarism

Hypotheses generation
Transliteration and transcription

Idea:

Language speakers tend to preserve phonetic and orthographic properties of the borrowed word

1) Transliteration

2) Transcription

Transliteration

Speaker has internal intuition about the writing system of the foreign language

Corpus: Dyakov dictionary of anglicisms + manually created set

Method: Statistical

1) separate on syllables
2) generate bigrams and trigrams of syllables
3) weight
4) result - max possible combination of syllables

Transcription

Idea: Speaker is supposed to preserve word's phonation while writing English word with Cyrillic script

Source: In-vocabulary lexis with pre-defined transcriptions — Cambridge Advanced Learner’s Dictionary, Cambridge Academic Content Dictionary and Cambridge Business English

Method: context-dependent grammar + additional rules, based on practical transcription of English named entities, proposed by Gilyarevsky.

Hypotheses generation
Transliteration and transcription

List of possible hypotheses for English words, how they could be written in Russian

EN word	EN-RU	TR-RU	RU-EN	RU word
football	футбол(0)	футбол(0)	footbol(2)	футбол
brainstorm	браинсторм(1.5)	брэйнстом(1)	bryeynshtorm(3)	брейншторм
fashion	фашион(2)	фэшэн(1)	feshn(3)	фэшн

Table 1. *Transliteration/Transcription hypotheses with Levenshtein distance.*

Hypotheses generation
Word root extraction

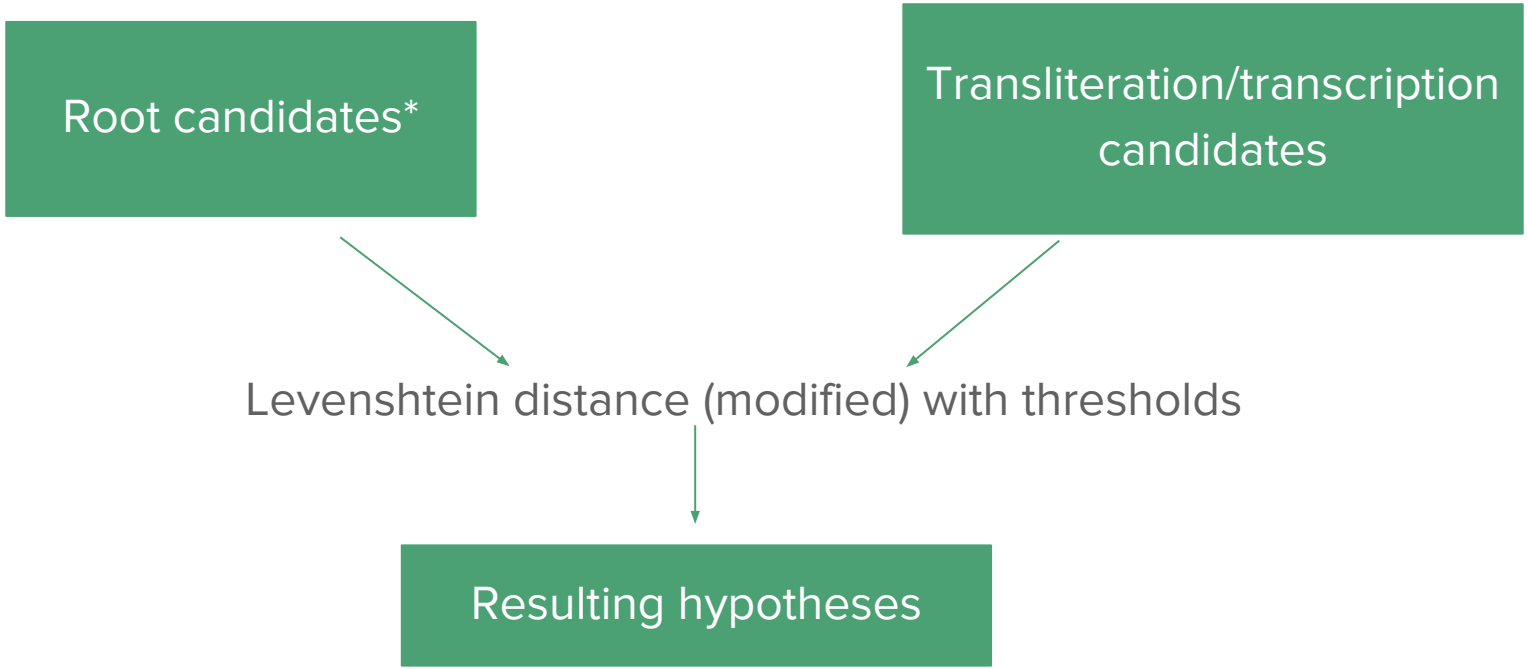
Idea: Frequent type of anglicisms — Russian word contains English root plus Russian affixes

Method: RNN based root extraction method

Model:

- trained on 97,000 pairs (*normal form — root*)
- roots were extracted from WikiDictionary
- non-vocabulary words with frequency >= 30

Hypotheses generation. Levenshtein comparison



**for compounds words as well (the two roots of the compounds were also checked)*

Hypotheses filtering

Idea: many anglicisms are used in the same context in both English and Russian spelling in social network texts

Method: distributive semantics

Algorithm: SkipGram

- Resulting hypothesis candidates in Cyrillic → Model (top 100 similar words)
- If our English hypothesis in SkipGram set → Word is anglicism!

Hypotheses filtering

Problem: Many anglicisms are rarely used by Russian speakers in original (English) spelling

Solution: Translation and context search with CBOW model

For each hypotheses all contexts, containing 5 words left and 5 words right the hypotheses are translated.

If top 100 most relevant English words for each context contains English analogue of the hypotheses in more than 50% cases, we consider the hypotheses to be proved.

Evaluation. Comparison with test set

The dictionary of anglicisms by A.I.Dyakov:

- is available online since 2014
- contains about 15,000 lexical items (1000 collocations)
- has a wide range of living spheres (economics, IT, marketing, etc.)
- involves loan words from spoken language, various slangs, professional jargons and profanities

Evaluation

Match the resulting list with manually created test set

Only 4321 words from Dyakov dictionary are available in model (LiveJournal).

Available in SkipGram model 2417 words

Manually annotated all hypotheses with LD < 1, missed at the Dyakov dictionary

1146 anglicisms were found by the method (863 + 283)

F-measure	Precision	Recall
0.38	0.84	0.24

Conclusion

1. Method is fully automated, it works and finds anglicisms!
2. About 1146 of 4300 words from manual dictionary was found
3. Language transfers. Method catches borrowings not only from English to Russian, but vise versa as well (Ex.: *pogrom, vodka*.)
4. The results are valuable for theoretical researches and can be applied in practical systems.

Future work

Improve transliteration/transcription. Reduce hypotheses:

- Russian diphthongs are not the same in English. Some English cases can not be transferred in Russian in this way.
- Morphological dependencies (*song - сонг,* but *running - раннин*)

Interesting anglicisms cases:

- English loan words borrowed from third languages (Ex.: *Sheikh*)
- Difficult cases such as homonymy (Ex.: *пост, док*)

Future work

- Comparison study of anglicisms in different languages (processes of language contacts — German vs. Russian vs. French examples)
- Processes of borrowings depending on the time of borrowing (Ex.: *борда* or *борд*)
- Usage of another corpora (Twitter, VKontakte)
- Online service with lists of new anglicisms

Thank you for attention!