# Correcting Keyboard Layout Errors and Homoglyphs in Search Queries

Derek Barnes, Mahesh Joshi, and Hassan Sawaf

{debarnes, mahesh.joshi, hsawaf}@ebay.com

## Two Key Challenges with Query Translation for Cross-border e-Commerce

8% of Russian queries with empty search results contain either a KLE or a homoglyph

### Keyboard Layout Errors (KLEs)

- Tokens entered in an unintended keyboard layout
- “чехол шзфв” instead of “чехол ipad” (“шзфв” maps to “ipad”, see mapping below)
- KLE mapping sample:

| English Layout | Russian Layout |
|---|---|
| i | ш |
| p | з |
| a | ф |
| d | в |
| … | … |

### Homoglyphs

- Tokens containing visually similar characters from different keyboard layouts
- “case” (underlined letters Cyrillic)
- Homoglyph mapping sample:

| English | Cyrillic |
|---|---|
| c | с |
| a | а |
| K | К |
| m | м |
| … | … |

## Sequence Labeling Approach

### Goal

Map KLEs and homoglyphs into ASCII prior to query translation

### Label Set

| Label | Meaning | Examples |
|---|---|---|
| R | Intended Russian token | Чехол, айфон |
| E | Intended English token | Case, iPhone |
| K | Keyboard Layout Error | сфк (car)<br>ыфьыгтп (samsung) |
| H | Homoglyph | case, вмw, e5 |
| A | Ambiguous token, all characters valid in both keyboard layouts | 123, &, / |

### Steps

1. Label each token in the query
2. Transform tokens that are KLEs or homoglyphs using the deterministic mapping (sample shown above)
3. Translate
4. Search

### Features

| Feature Set | Description | Examples |
|---|---|---|
| Language model features | 5-gram character-level language models, trained on:<br>• general English and Russian corpora<br>• list of brand names<br>• Russian transliterations for proper names | • “Чехол” scores high in Russian LMs, but its KLE mapping “xt[jk” scores low in English LMs<br>• “сфыу” scores low in Russian LMs, but its KLE mapping “case” scores high in English LMs |
| Dictionary features | Does token exist in one of the following lexicons:<br>• Unix English dictionary (~480K words)<br>• A curated list of 58K brands<br>• A lexicon of ~1.8M words from 10M eBay EN titles | Features fire for the surface form of a token, and its KLE and homoglyph mappings if they are present in the lexicon |
| Token class | Standard features, such as case, shape, token class, position, etc.<br>*Note*: Lexical features such as word id did not help | • is token all UPPERCASE?<br>• does token contain numbers? |

### Algorithms

First order conditional Markov models that use logistic regression and Random Forests as base classifiers

### Rule-based Baseline

1. Any token among a list of 101 Russian stopwords tagged as ‘R’
2. Fully ASCII tokens labeled as ‘A’ if all its characters are common to both keyboard layouts, else labeled ‘E’
3. Cyrillic tokens labeled as ‘K’/‘H’ if their KLE/homoglyph mappings are found in any of the lexicons (ties broken randomly), else labeled ‘R’

*Note*: ‘E’ and ‘A’ tags can be perfectly labeled, since we only consider the Russian to English direction

## Results and Analysis

### Dataset

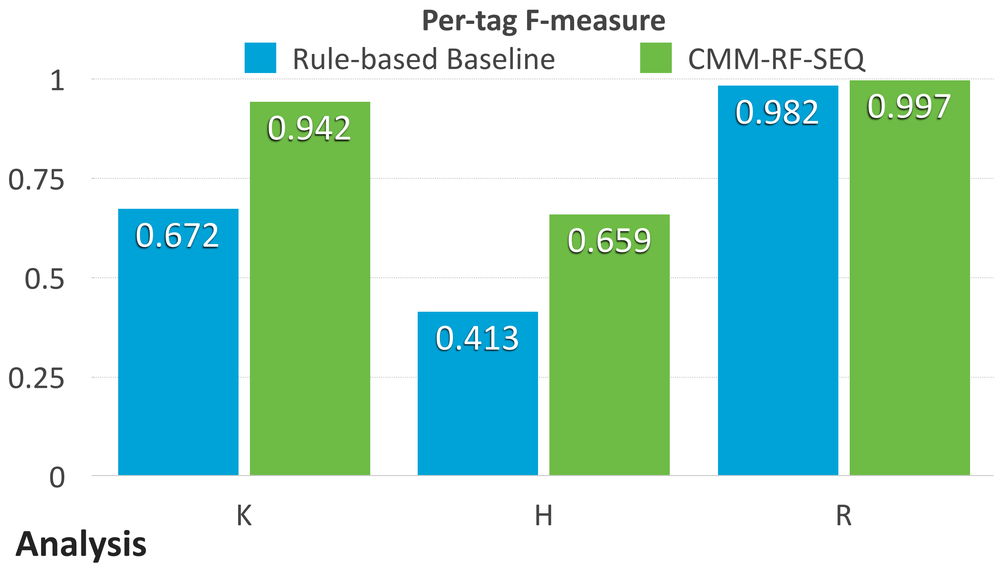
- *Training Set*: 6472 human- labeled query examples (17,239 tokens).
- *Test Set*: 2500 Russian/English queries (8,357 tokens) randomly selected from queries with null search results.
- Tokens labeled by a team of Russian language specialists

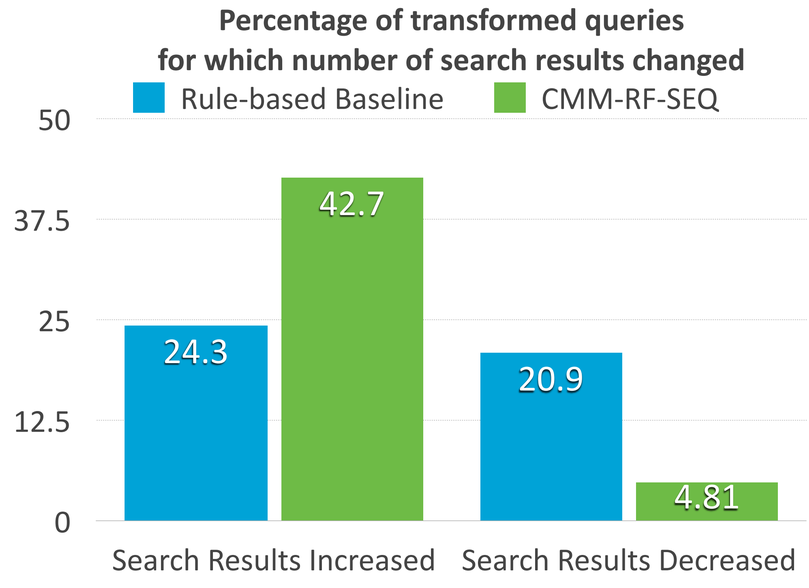
### Impact on Search Results

- 100K randomly selected Russian/English queries
- Transformed using baseline, and using our CMM-RF model
- Translated and searched

### Results

CMM-RF-SEQ is a first order conditional Markov model based on Random Forests




Baseline transformed: 12,661 of 100K queries

CMM-RF-SEQ transformed: 7,364 of 100K queries

### Analysis

- Character-level language models provide the most predictive power
- Short words and acronyms pose the greatest challenge
  - Russian car brand “ваз” maps across keyboard layouts to a common acronym “dfp” (Digital Flat Panel).
  - Russian words “муки” and “рук” map by chance to English words “verb” and “her”.
  - In Cyrillic query “БМВ е46”, “е46” can be interpreted either as a homoglyph or a KLE for ASCII “t46”