Saint Petersburg
State University
www.spbu.ru

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Alina Dubatovka / SPbSU

Yurii Kurochkin / Yandex

Elena Mikhailova / SPbSU

Dialogue 2016, Moscow, June 1-4, 2016

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Goals

- ~~Automatic extraction of sentiment words~~
- Automatic polarity detection
- Unsupervised

Methodology

- Hatzivassilogloum, McKeown 1997
  - "Tasty and healthy Breakfast“
  - "Cheap but nice hotel“
- The better the node is connected with other "positive" nodes and the worse with the "negative", the more positive it is

Graph builder

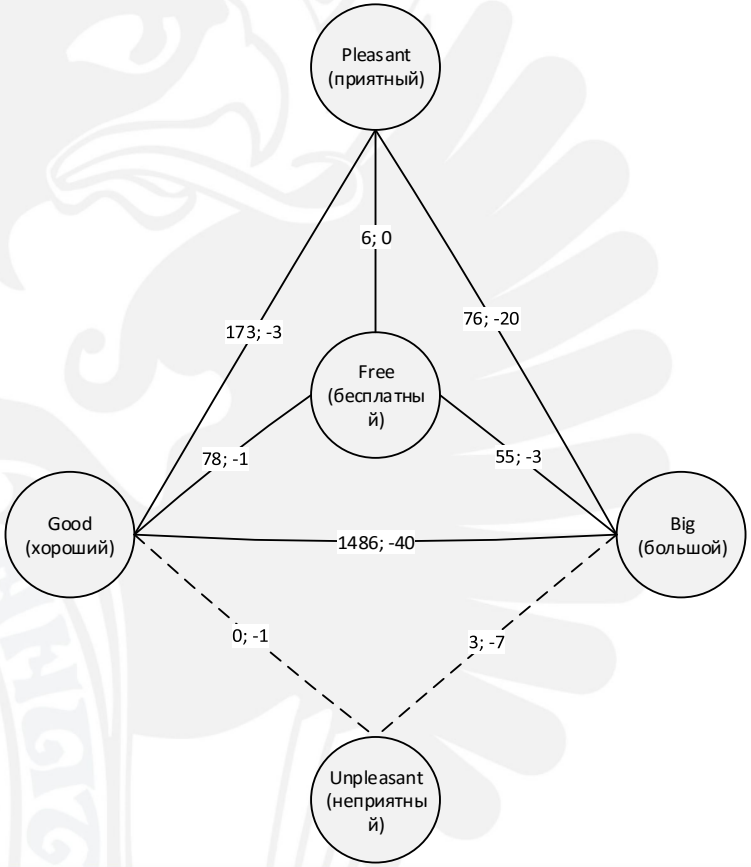
- $(\mathrm{ADV}\,|\mathrm{NEG}) * \mathrm{ADJ}(,?\,(\mathrm{AND}\,|\mathrm{BUT})?\,(\mathrm{ADV}\,|\mathrm{NEG}) * \mathrm{ADJ}) +$
- AND – conjunction "and“
- BUT – one of adversative conjunctions ("but", "instead", "however”, “nevertheless ")
- NEG – negation
- ADV – an adverb of measure and degree ("very", "quite", "too", "completely")
- ADJ – adjective

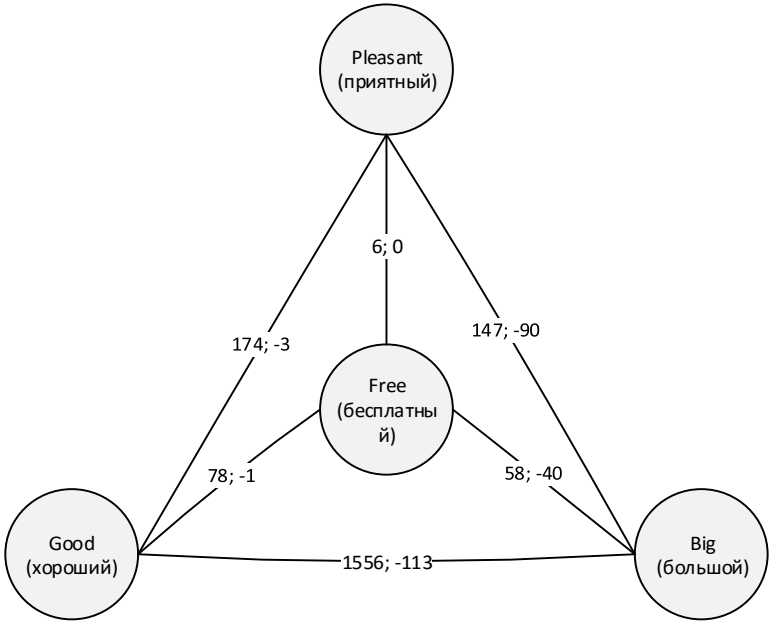
Example

- "*Tasty, plentiful but not very varied and expensive breakfast*"
- positive links: (*tasty, plentiful*), (*tasty, varied*), (*plentiful and varied*)
- negative links: (*tasty, expensive*), (*plentiful, expensive*), (*varied, expensive*).

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Particle “not” and prefix “un-”





Graph Analyzer

- Initialization
- Weight of the graph edges
  - $weight(word_1, word_2) = \#(word_1 AND\ word_2) - K * \#(word_1 BUT\ word_2)$
- Distance to the final set
  - The heaviest edge
  - The sum of the weights of edges

Description of experiments

- 259023 depersonalized unlabeled reviews
- Dataset size – 660 Mb
- Hotel domain
- Texts by real users
  - Misspellings
  - Grammatical errors
  - Informal words
  - unrelated information concerning flight, excursions, places of interest *etc*

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

“Large” dictionaries

	Positive	Negative	Neutral	Total
Algorithm without removing the "un-" prefix	5252	2815	-	8067
Algorithm after removing the "un-" prefix	4936	2695	-	7631
“Large” dictionary	1948	1946	4951	8845

AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

“Small” dictionaries

	Positive dictionary	Negative dictionary	Total
“Manual” dictionary	173	127	300
Algorithm without “un-” prefix removing	164	74	238
Algorithm with “un-” prefix removing	163	83	246

Results without removing the "un-" prefix

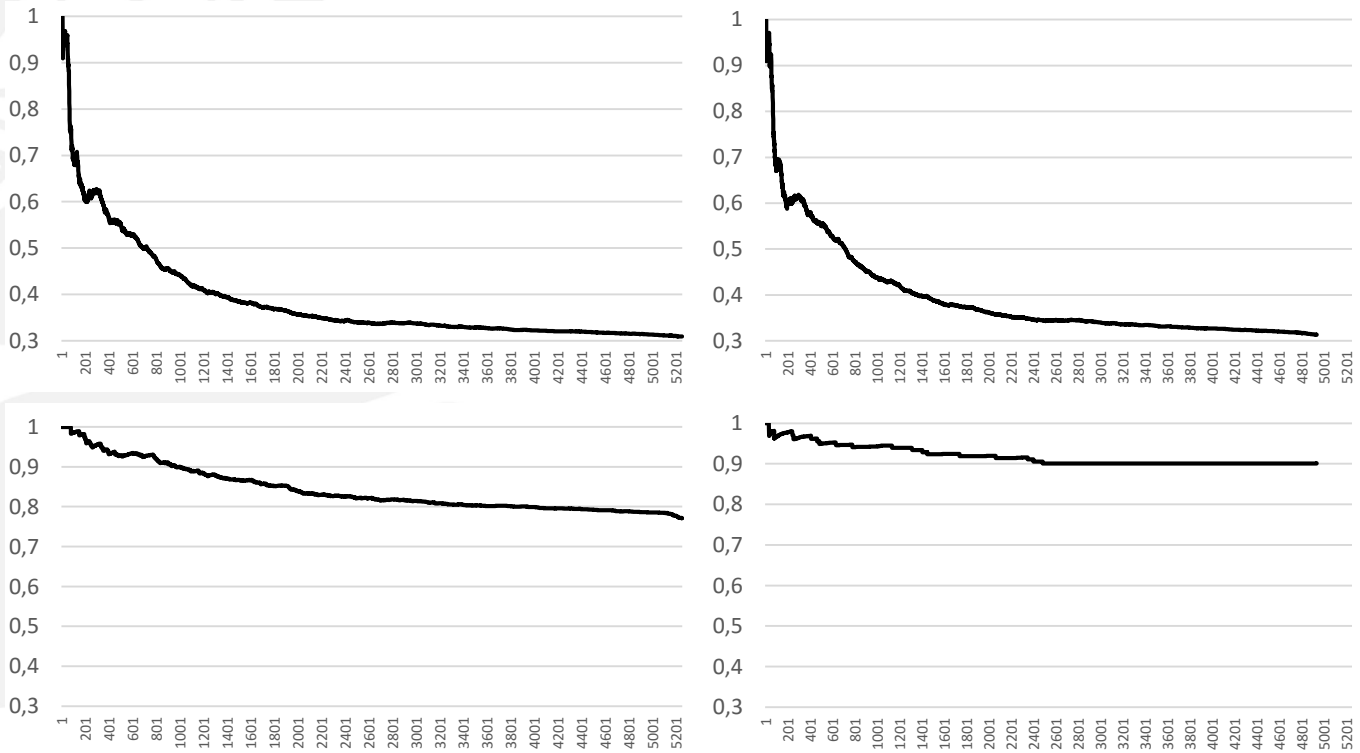
Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.806	0.684	0.754
Precision	0.309	0.521	0.381
Precision without neutral words	0.77	0.827	0.796
F_1-measure	0.447	0.591	0.506
F_1-measure without neutral words	0.788	0.749	0.774

Results after removing the "un-" prefix

Metric	Positive dictionary	Negative dictionary	Total dictionary
Recall	0.793	0.683	0.746
Precision	0.314	0.502	0.38
Precision without neutral words	0.779	0.82	0.799
F_1-measure	0.45	0.579	0.504
F_1-measure without neutral words	0.786	0.745	0.772

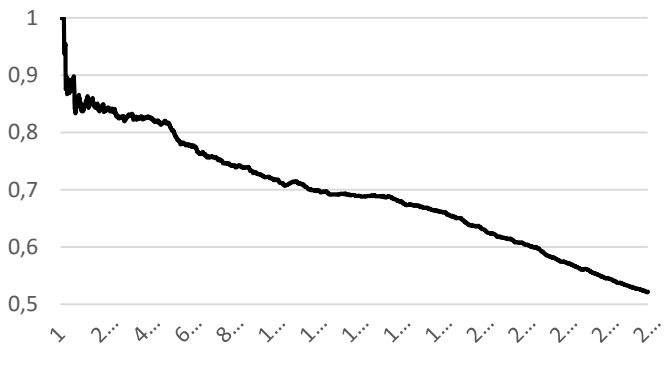
AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

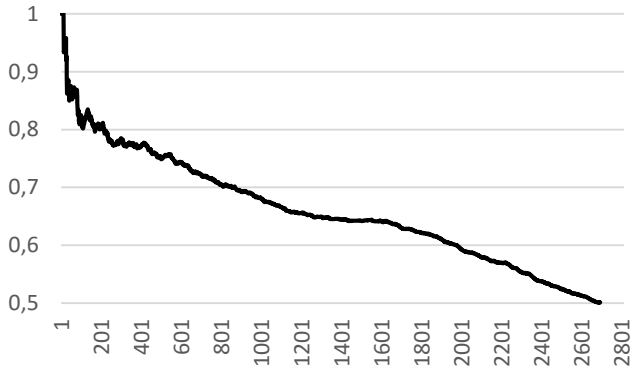
Precision@n for positive dictionary

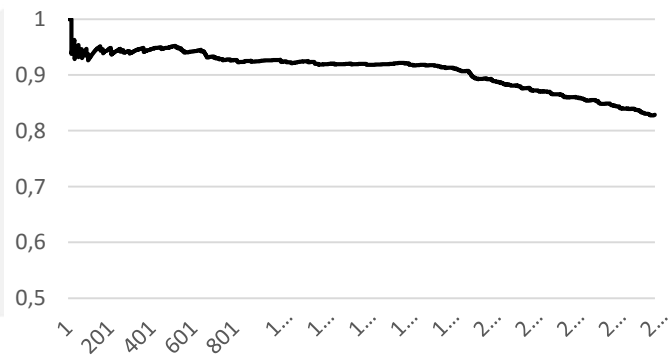


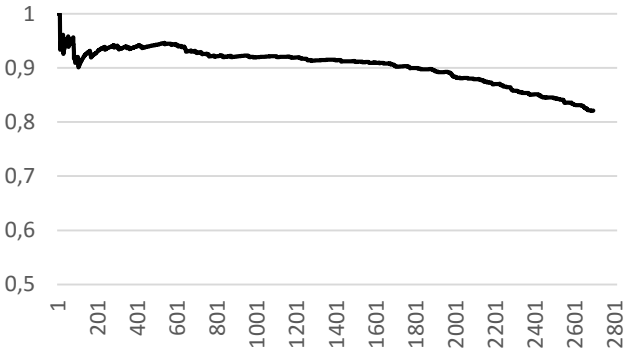
AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Precision@n for negative dictionary



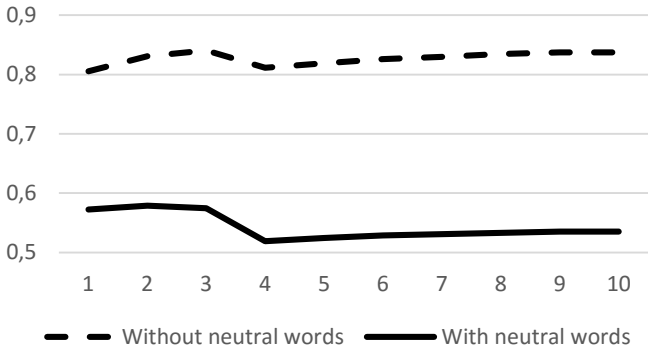


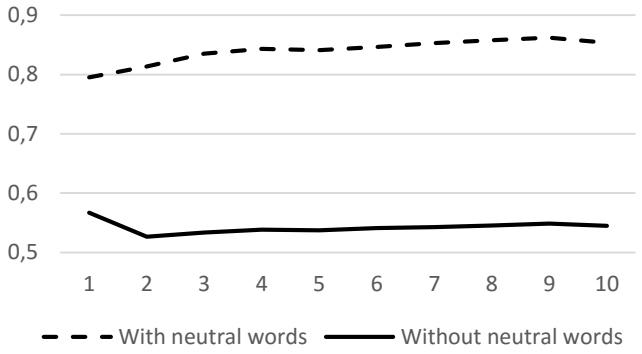


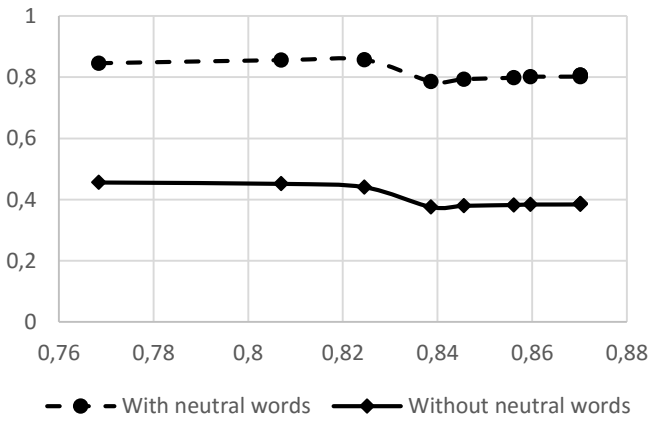


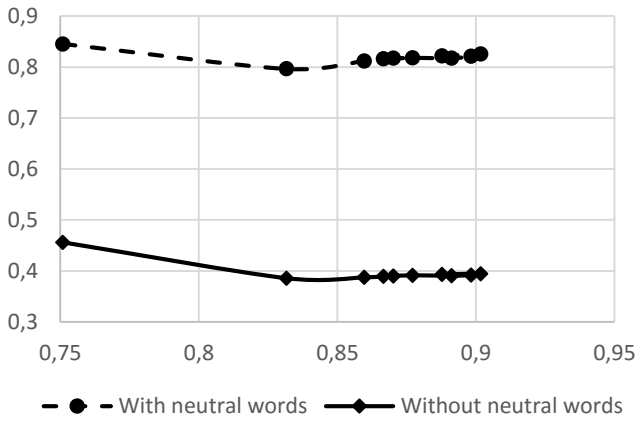
AUTOMATIC GENERATION OF THE DOMAIN-SPECIFIC SENTIMENT RUSSIAN DICTIONARIES

Dependence on K









Thanks!

St. Petersburg University

spbu.ru