Interactive Coding of Responses to Open-Ended Questions in Russian

Nikita Senderovich   Archil Maysuradze

Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University

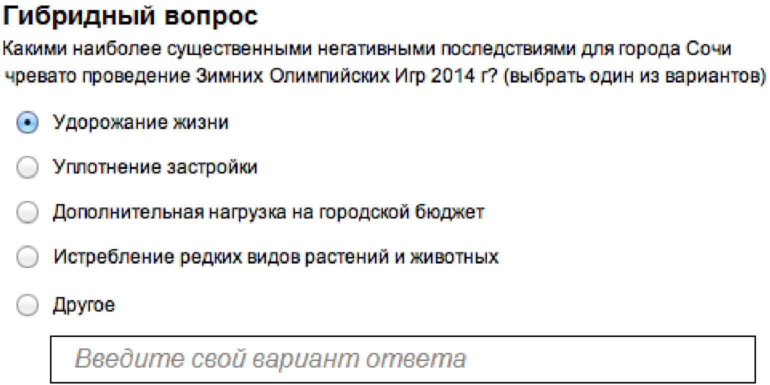
KESW, 2015

Agenda

1. The Task of Coding Open-Ended Questions
2. Concept of the Interactive System
3. Data Processing Steps
4. The Developed System

Types of Questions in Surveys

- closed-ended questions — choice from fixed set of answers
- **open-ended questions** — response in respondent's own words
- hybrid questions:



Гибридный вопрос

Какими наиболее существенными негативными последствиями для города Сочи чревато проведение Зимних Олимпийских Игр 2014 г? (выбрать один из вариантов)

● Удорожание жизни

○ Уплотнение застройки

○ Дополнительная нагрузка на городской бюджет

○ Истребление редких видов растений и животных

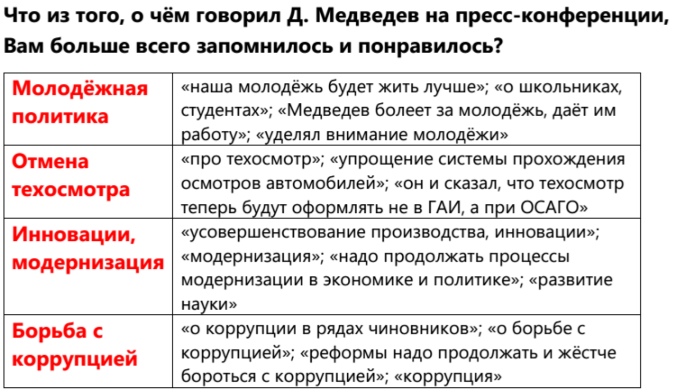
○ Другое

Введите свой вариант ответа

General Approach to Coding Task

- extraction of ideas from the answers
- creation of the codebook
- assigning one or more codes to each answer

Example of Coding Result:



Что из того, о чём говорил Д. Медведев на пресс-конференции, Вам больше всего запомнилось и понравилось?

Молодёжная политика	«наша молодёжь будет жить лучше»; «о школьниках, студентах»; «Медведев болеет за молодёжь, даёт им работу»; «уделял внимание молодёжи»
Отмена техосмотра	«про техосмотр»; «упрощение системы прохождения осмотров автомобилей»; «он и сказал, что техосмотр теперь будут оформлять не в ГАИ, а при ОСАГО»
Инновации, модернизация	«усовершенствование производства, инновации»; «модернизация»; «надо продолжать процессы модернизации в экономике и политике»; «развитие науки»
Борьба с коррупцией	«о коррупции в рядах чиновников»; «о борьбе с коррупцией»; «реформы надо продолжать и жёстче бороться с коррупцией»; «коррупция»

Manual Coding and Its Problems

Typical Process:

- senior analyst creates the codebook
- group of analysts performs the coding

Problems:

- laboriousness
  - consistent codebook modification
  - intercoder agreements
- subjectivity

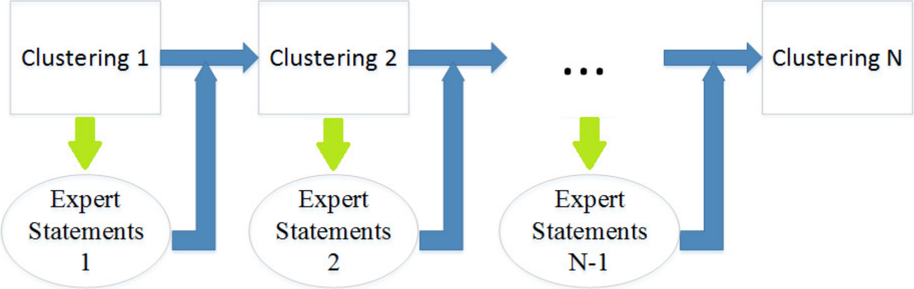
No proven industry standard for automated coding exists.

Research Goals

Propose and study the coding process with the following properties:

- work in group
- coding result consistent with opinion of members of the group
- inductive approach to building coding scheme

Interactive Clustering



Domain-Oriented Set of User Statements

1. select the subset of responses
2. attach selected subset to existing cluster
3. attach selected subset to new cluster
4. detach responses of selected subset from clusters they are attached to
5. withdraw selected subset from consideration
6. complete the formation of cluster
7. continue the formation of cluster
8. remove the cluster

Theorem

Statements 1, 3, 5, 8 allow to achieve arbitrary clustering.

Cooperative Workflow

- group of analysts make statements through web-interface in real time

Cooperative Workflow

- group of analysts make statements through web-interface in real time
- consistent current state is presented at all times

Cooperative Workflow

- group of analysts make statements through web-interface in real time
- consistent current state is presented at all times
- the result is validated online

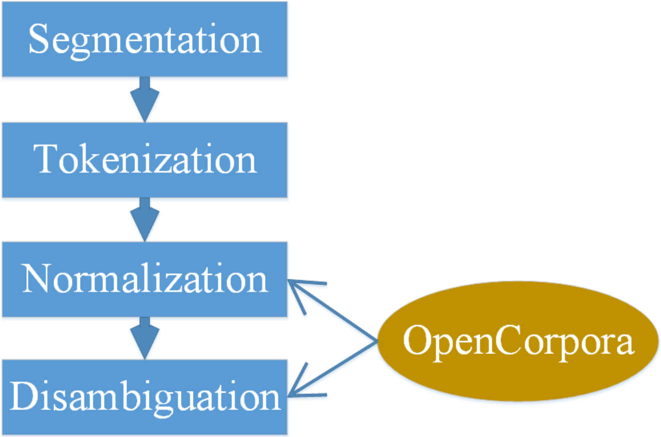
Cooperative Workflow

- group of analysts make statements through web-interface in real time
- consistent current state is presented at all times
- the result is validated online
- matters of opinion are detected and agreement is achieved

Cooperative Workflow

- group of analysts make statements through web-interface in real time
- consistent current state is presented at all times
- the result is validated online
- matters of opinion are detected and agreement is achieved
- objectivity is increasing

Natural Language Processing



Vector Space Model

W — dictionary of all terms

D — set of responses

n_{dw} — number of occurences of term w in document d

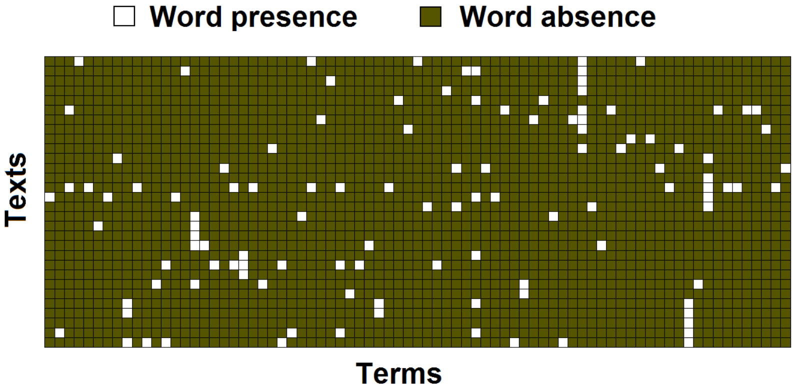
$d = [f_1^d, \dots, f_{|W|}^d]^T$, where $f_w^d = [n_{dw} > 0]$, $d \in \mathbb{R}_+^{|W|}$

Standard similarity function — cosine measure:

$$s(u, v) = (u, v), \qquad ||u|| = ||v|| = 1.$$

Source Data Sparseness Problem

Part of Typical Term-Document Matrix:



Source Data Sparseness Problem

Problem: lack of common context information

Sources of additional information:

- semantic graphs and nets
- expert opinion

Semantic smoothing method:

Russian Thesaurus (RuThes) $\longrightarrow$ semantic proximity matrix $P \longrightarrow$

$$\text{similarity function } s'(u,v) = \frac{(Lu, Lv)}{||Lu||||Lv||},\ L^T L = P \longrightarrow$$

$$\text{distance function } d(u,v)$$

Text Clustering

- **hierarchical algorithms** (agglomerative and divisive):
  - Single linkage clustering: $d(C_i, C_j) = \min\limits_{x \in C_i, y \in C_j} d(x, y)$
  - UPGMA clustering: $d(C_i, C_j) = \dfrac{1}{|C_i||C_j|} \sum\limits_{x \in C_i} \sum\limits_{y \in C_j} d(x, y)$
  - DIANA clustering
- **spherical k-Means**
  - applying smoothing: $x_j = \dfrac{Px_j}{||Px_j||}$
  - formulating an optimization problem:
    $$\sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j, c_i) \longrightarrow \max_{\boldsymbol{c}, \boldsymbol{r}}, \;\; ||c_i|| = 1$$
  - iterative solution:
    $$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \qquad c_i = \frac{\mu_i}{||\mu_i||}, \qquad r_j = \mathrm{argmax}_i (x_j, c_i)$$

Text Clustering: Experimental Results

The quality is measured using *F*-measure.

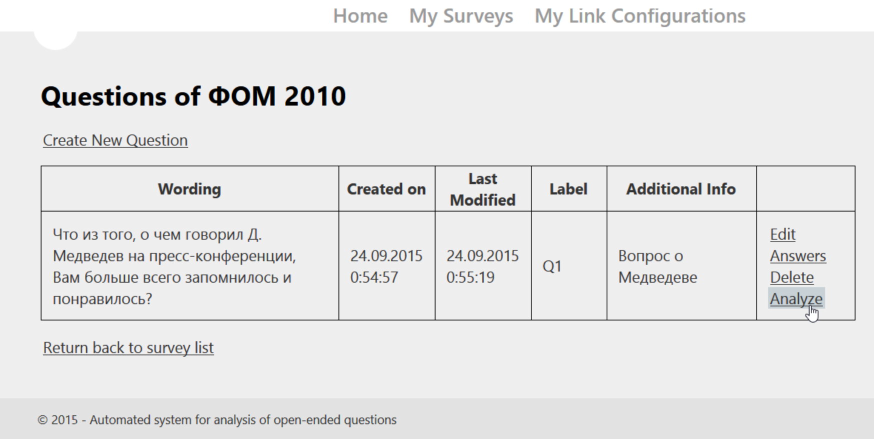
Method	**UPGMA**		**SL**		**DIANA**		**SKM**	
Smoothing	no	yes	no	yes	no	yes	no	yes
M1	1.00	**1.00**	1.00	**1.00**	1.00	**1.00**	1.00	**1.00**
M2	0.89	0.90	0.78	0.87	0.66	0.93	1.00	**1.00**
M3	0.61	0.75	0.35	0.53	0.48	0.56	0.79	**0.81**
M4	0.61	0.67	0.35	0.47	0.48	0.57	0.79	**0.80**

Semantic smoothing significantly enhances clustering performance

Spherical K-means shows the best results

In spherical K-means user statements can be easily formalized

The Developed System



Questions of ФОМ 2010

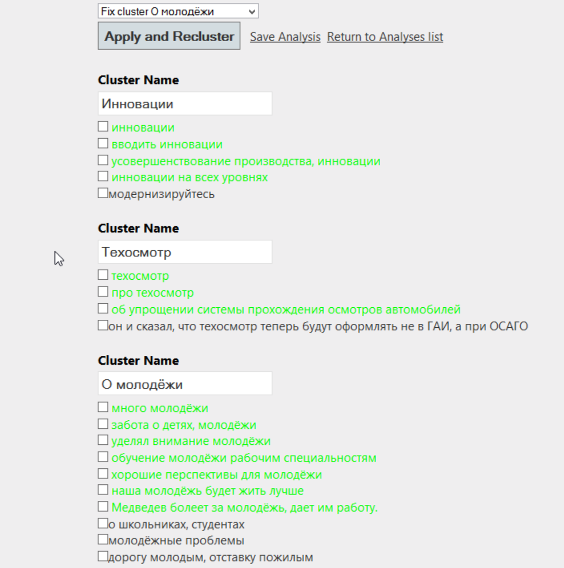
Create New Question

Wording	Created on	Last Modified	Label	Additional Info	
Что из того, о чем говорил Д. Медведев на пресс-конференции, Вам больше всего запомнилось и понравилось?	24.09.2015 0:54:57	24.09.2015 0:55:19	Q1	Вопрос о Медведеве	Edit Answers Delete Analyze

Return back to survey list

© 2015 - Automated system for analysis of open-ended questions

The Developed System



Cluster Name

Инновации

- □ инновации
- □ вводить инновации
- □ усовершенствование производства, инновации
- □ инновации на всех уровнях
- □модернизируйтесь

Cluster Name

Техосмотр

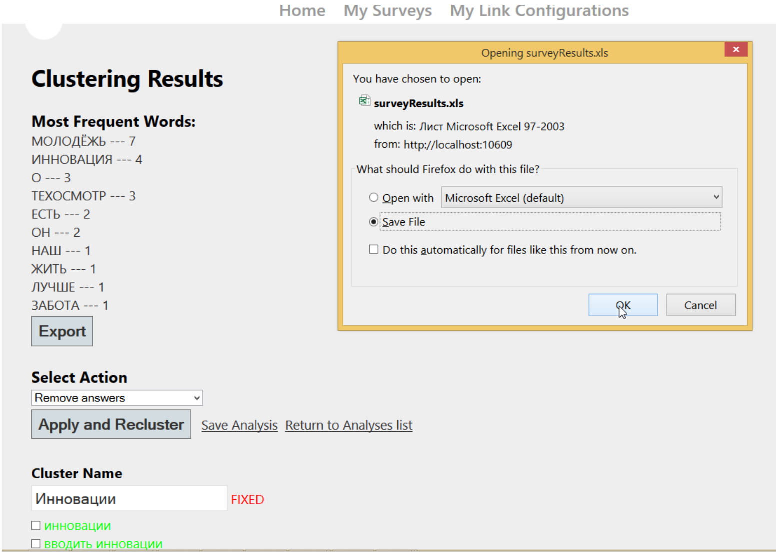
- □ техосмотр
- □ про техосмотр
- □ об упрощении системы прохождения осмотров автомобилей
- □он и сказал, что техосмотр теперь будут оформлять не в ГАИ, а при ОСАГО

Cluster Name

О молодёжи

- □ много молодёжи
- □ забота о детях, молодёжи
- □ уделял внимание молодёжи
- □ обучение молодёжи рабочим специальностям
- □ хорошие перспективы для молодёжи
- □ наша молодёжь будет жить лучше
- □ Медведев болеет за молодёжь, дает им работу.
- □о школьниках, студентах
- □молодёжные проблемы
- □дорогу молодым, отставку пожилым

The Developed System



Clustering Results

Most Frequent Words:

МОЛОДЁЖЬ --- 7
ИННОВАЦИЯ --- 4
О --- 3
ТЕХОСМОТР --- 3
ЕСТЬ --- 2
ОН --- 2
НАШ --- 1
ЖИТЬ --- 1
ЛУЧШЕ --- 1
ЗАБОТА --- 1

Select Action

Remove answers

Cluster Name

Инновации FIXED

□ инновации
□ вводить инновации

Opening surveyResults.xls

You have chosen to open:

surveyResults.xls

which is: Лист Microsoft Excel 97-2003
from: http://localhost:10609

What should Firefox do with this file?

○ Open with Microsoft Excel (default)
● Save File
□ Do this automatically for files like this from now on.

Expert Coding: Experimental Results

Responses to three open-ended survey questions were coded using the implemented web service.

Data Set	Responses	Clicks	Iterations	Clusters
Q1	43	17	4	3
Q2	125	49	15	10
Q3	727	130	27	17

Number of Clicks $<$ Number of coded responses

Conclusion

- The coding process with properties corresponding to domain requirements is proposed.
- Usage of machine learning and text mining techniques as an auxiliary instrument in coder's work allowed to achieve effort minimization.
- Implementation of web interface for open-ended coding allows to use cooperative methodology of coding, which enhances the quality of results.

References I

Loukashevich N.V.
Thesauri in problems of information retrieval.
Moscow University Printing House, 2011.

Varlamov M. I., Korshunov A. V.
Computing semantic similarity of concepts using shortest paths in Wikipedia link graph.
JMLDA, 1107-1125, 2014.

Boyarsky K.K., Kanevsky E.A., Saganenko G.I.
On the issue of automatic text classification.
Economic-mathematical studies: mathematical models and information technology, 253-273, 2009.