TESTING RULES FOR SENTIMENT ANALYSIS SYSTEM

E. Kuznetsova (knnika@yandex.ru)
GK "Geostream",
N.Loukachevitch (louk_nat@mail.ru), I. Chetviorkin (ilia2010@yandex.ru)
Lomonosov Moscow State University

Sentiment classification of texts

- Machine learning approaches
  - Need training collections
  - If there are enough training collections then performance better than..

- Knowledge-based approaches
  - Sentiment lexicons+rules
  - Require manual labor but do not require training collections

ROMIP-2012 sentiment evaluation

- News quotation classification
  - 3 classes (positive, negative, neutral)
- Example of negative quotation:
- *По мнению эксперта, глава белорусского государства больше всего **боится (afraid of),** что страну все-таки **лишат права** (**deprive the right**) провести чемпионат мира по хоккею в 2014 году*.

- News quotations in a broad domain:
  - difficult to create a good training collection

Outline

- Polarnik – lexicon-based sentiment classification system
- Participation of Polarnik in news quotation evaluation at ROMIP 2012
- Improvements in Polarnik performance based on a new rule set
  - Testing additional rules

Polarnik system

- Dictionary+ simple rules
- Dictionary of words and expressions (constructed semi-automatically)
  - Positive lexicon (7 thousand)
    - Бесподобный, безболезненный
  - Negative lexicon (15 thousand)
    - Бесправный, бесполезный
  - Word-operators (*не, очень, уменьшить, увеличить*) - 140
  - Stop-expressions («Фонд эффективной политики» - *Foundation of effective politics*) - 250

Semi-automatic population of sentiment lexicon

- Initial point – sentiment lexicon in movie domain
- Extraction of news articles:
  - at least three sentiment words from movie sentiment lexicon
  - given share of sentiment words

- Two news collections:
  - Coll-news – 2mln documents – initial collection
  - Coll-sent – collection with presumably high share of sentiment words
  - Extraction of words with $df_{sent}>100$

Semi-automatic population of sentiment lexicon - 2

- Extraction and ordering words from Coll-sent using:

$$Weirdness = \frac{P_s(w)}{P_g(w)}$$

$P_s(w)$ – probability of the word appearance in documents of *Coll_sent* sub-collection,
$P_g(w)$ – probability of the word appearance in documents of *Coll_news* collection.

- The first 10 thousand words were looked through by experts – 30% of new sentiment words
- Last stage: testing lexicon coverage in sentiment analysis of large news articles (analytics)

Rules in Polarnik

- Rules
  - Summation of sentiment scores of words and expressions
    - Interesting (+1) + nice (+1)=2
    - Interesting (+1) + sad (-1)=0
  - Multiplying operator score and sentiment score
    - Not (-1) *interesting(+1)=-1
    - Very (2) *interesting(+2)=2

Results in quotation evaluation (РОМИП-2012)

Run_ID	*Macro_P, %*	*Macro_R,%*	*Macro_F1,%*
Polarnik	**62.6**	**61.6**	**62.1**
xxx-11	60.6	57.9	59.2
xxx-15	56.3	56.0	56.2

Analysis of Polarnik errors in ROMIP train collection

- 40 quotations with differences between automatic and manual sentiment scores were taken
- - 16 quotations – lack of sentiment words and expressions in the system lexicon (40%)
  - Among then 13 quotations – lack of sentiment expressions or stop expressions
- - 5 quotations: disagreement with experts (12.5%)
- - (!) classification of 4 quotations (10%) could be improved using additional set of rules

Examples of quotations, which can be improved by additional rules

- Он заявил, что речь идет о **прискорбном недоразумении**, ведь он всегда считал, что (thought that) литература и **искусство** должны служить **морали**
  - **System: 0**
  - **Expert: -**
- Секретарь президиума генсовета «Единой России», зампредседателя Госдумы Сергей Неверов в субботу заявил, что партия **не боится раскола** (do not afraid of split) в связи с появлением в ней разных идеологических платформ
  - System: 0
  - Expert: +

Study of additional rules impact

- Add new rules and do not change lexicon
- Tune rules on training set of quotations
- Test the whole system
  - On training set
  - On testing set

Rule set 1 (algo)

- If an operator word is a part of a longer stop or sentiment expression, it does not act as an operator;
- If a group of operators appears together, their scores are multiplied;
- If there is unknown hyphenated word appeared in a text fragment, it is divided in two words and their scores are considered separately;
- If there is a sentiment word sequence, and a negative word appears among them then the score of the whole sequence becomes negative, otherwise positive;
- An operator is applied to the resulting score of a group of sentiment words.

Rule set 2 (rules)

- Account for “irrealis markers”
- Sentiment score of sentence (sentence fragment) should be reduced if:
  - there is a question mark in a sentence
  - there is *если (if)* in a fragment;
  - there is *ли* particle in a fragment, and there is no such words as ч*уть*/*то*/*вряд*/*видишь*/*видите*/*мало*/*едва*/*что* just before *ли*;
  - there is *бы* particle in a fragment

- Rules are based on real examples in training collection

Results of new rules on quotation training collection

	Macro_P%	Macro_R%	Macro_F1%	Accuracy%
Baseline	60.9	61.0	60.9	60.5
Baseline+ rules	61.1	61.3	61.2	60.9
Baseline+ algo	61.4	61.5	61.5	61.4
Full compo-sition	61.5	61.6	61.6	61.5

Changes in correct quotations

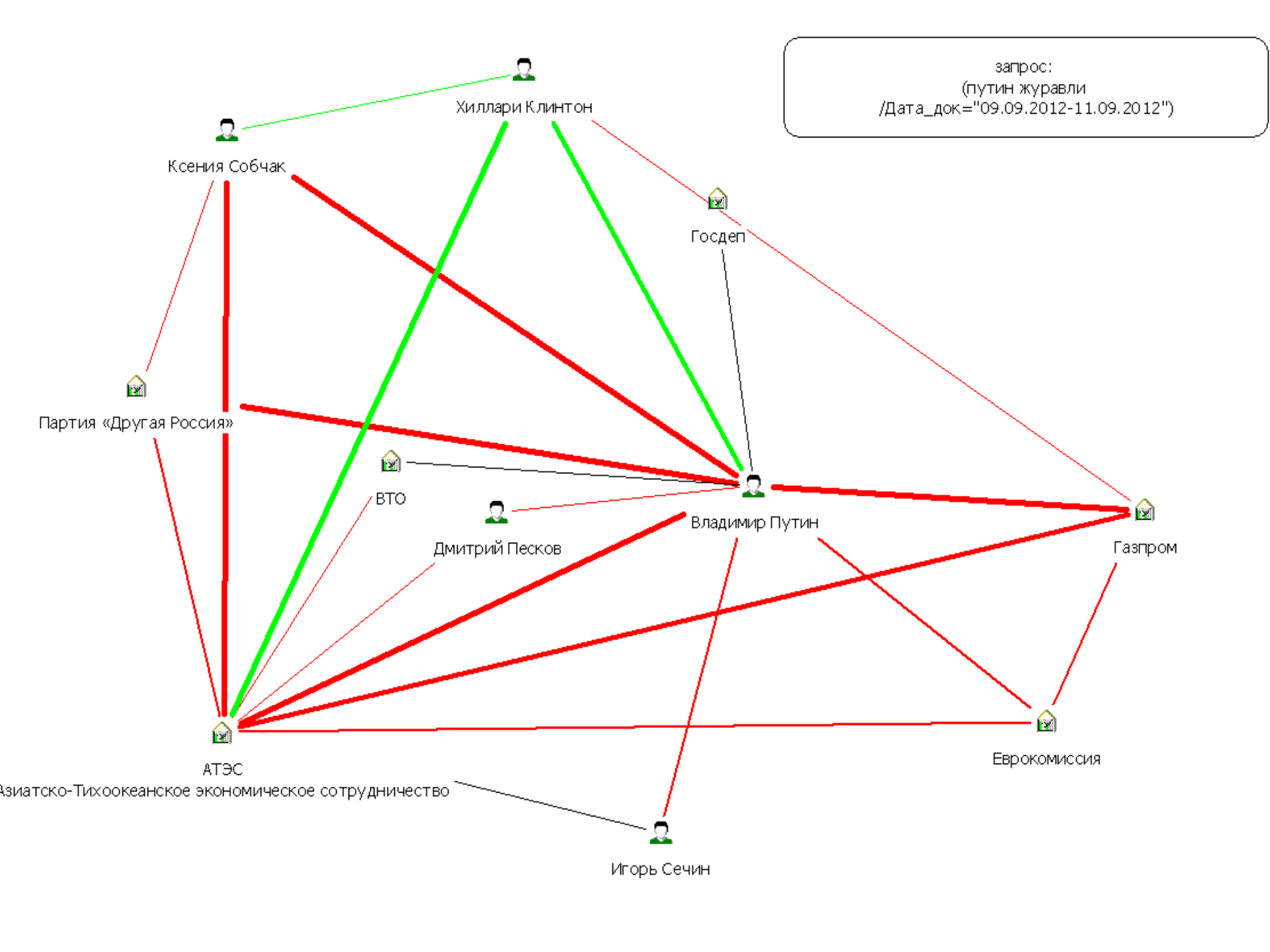
	Number of qoutations changed to correct class	Number of quotations changed to incorrect class	Growth of correct qoutations
Baseline	-	-	-
Baseline+ rules	20	7	13
Baseline+ algo	53	21	32
Full compo-sition	60	22	38

Results of new rules on quotation test collection

	Macro_P%	Macro_R%	Macro_F1%	Accuracy%
Baseline	62.6	61.6	62.1	61.6
Baseline+ rules	62.8	61.9	62.3	61.9
Baseline+ algo	63.0	62.2	62.6	62.25
Full compo-sition	62.9	62.2	62.6	62.32

Polarnik in Information Retrieval System

- In every document positions of positive/negative sentiment expressions are indexed
- These information is loaded to information-retrieval index as word position
- We can ask sentiment-oriented queries


запрос:
(путин журавли
/Дата_док="09.09.2012-11.09.2012")
Хиллари Клинтон
Ксения Собчак
Госдеп
Партия «Другая Россия»
ВТО
Дмитрий Песков
Владимир Путин
Газпром
АТЭС
Еврокомиссия
Азиатско-Тихоокеанское экономическое сотрудничество
Игорь Сечин

Информация | Дерево | Кластеры | Лог построения рубрикатора

Отчет по запросу: мгу суперкомпьютеры /САНТИМЕНТ="+"

Формирование отчета: по кластерам
Рубрицирование по тезаурусу: НАУКА (Расширение: L) [K: 50]

АЭРОДИНАМИКА

(0.5) 10.08.2012 11:00:00 **Садовничий рассказал Путину про суперкомпьютер и космические спутники, которые запускает университет** **[Накануне.RU] #8520503#**
Садовничий рассказал Путину про суперкомпьютер и космические спутники, которые запускает университет Суперкомпьютер, работающий в МГУ, помог создать новые глазные капли и расшифровать американскую криптографию, сообщил ректор МГУ Виктор Садовничий на встрече с президентом России Владимиром Путиным, которая, как сообщили Накануне.RU в пресс-службе Кремля, состоялась в четверг. [51сп; Cluster: 5]

ЭЛЕКТРОНИКА

(0.67) 18.04.2012 14:14:00 **В МГУ появится новая магистерская программа по супервычислениям** **[0Российская газета - RG.RU] #5149389#**
В США, Германии, Франции, Китае вычислительные центры стали национальными. Сейчас наука на таком этапе развития, когда многие открытия невозможны без мощной вычислительной базы, - подчеркнул он на семинаре в МГУ, где ведущие ученые университета рассказывали о своих последних достижениях. Причем, все эти достижения были получены с помощью суперкомпьютера, которым владеет МГУ. [75сп; Cluster: 1]

(0.62) 26.09.2012 00:14:26 **МГУ имеет все возможности включиться в петафлопную гонку** **[0Независимая Газета] #10001590#**
Гордостью и украшением центра должен стать новый суперкомпьютер. Предполагается, что многие разработки будут вестись на средства инвесторов. Новый Ломоносовский корпус МГУ предназначен для ведущих научных коллективов, занимающихся перспективными исследованиями, и призван стать «технологическим поясом» университета. [75сп; Cluster: 2]

(0.52) 09.08.2012 21:14:27 **Ректор МГУ Виктор Садовничий сообщил Президенту о ключевых проектах в работе ВУЗа** **[0Первый Канал (ОРТ - видео)] #8506530#**
Ректор МГУ Виктор Садовничий сообщил Президенту о ключевых проектах в работе ВУЗа Важным направлением Виктор Садовничий назвал работу на суперкомпьютерах - этим заняты 600 научных коллективов. В частности, благодаря этим... [33сп; Cluster: 3]

(0.5) 10.08.2012 01:00:00 **Виктор Садовничий - Владимиру Путину: «Наш суперкомпьютер оказался в три раза мощнее, чем задумывали»** **[Комсомольская правда] #8510405#**
Виктор Садовничий - Владимиру Путину: «Наш суперкомпьютер оказался в три раза мощнее, чем задумывали» В четверг Владимир Путин провел встречу с ректором МГУ Виктором Садовничим, в ходе которой обсудил программу развития крупнейшего вуза страны, рассчитанную до 2020 года. [41сп; Cluster: 4]

(0.5) 28.03.2012 07:35:00 **Top50 самых мощных компьютеров СНГ** **[0Ferra.ru - Компьютерные новости] #4517591#**
Научно-исследовательский вычислительный центр МГУ имени М.В.Ломоносова и Межведомственный Суперкомпьютерный Центр РАН выпустили шестнадцатую редакцию списка Top50 самых мощных компьютеров СНГ. Ломоносов Шестнадцатая редакция списка продемонстрировала

Conclusion

- Large dictionaries of Polarnik system provided the best results in ROMIP quotation evaluation
- Without any changes to the sentiment lexicon we implemented an additional set of rules to take into account groups of opinion words and operators and irrealis markers.
- Using these new rules, the system performed better both on the train and test collections