Extracting Sentiment Attitudes from Analytical Texts

Natalia Loukachevitch

Lomonosov Moscow State University

Nicolay Rusnachenko

Bauman Technical University

Sentiment analysis: genres of documents

- A lot of studies are devoted to sentiment analysis of users' reviews or short posts in social networks (Twitter)
  - One source of opinion – one target (in most cases)

- News or analytical reports
  - Many sources of opinions and many targets:
    - Author
    - Quotations
    - Participants to each other
  - Positive and negative events
  - Many names, which are not sources or targets of opinions

Example

- As is apparent in **Washington**, there is no place for objectivity on the subject of **Russia**, irrespective of facts and events.
- (http://www.counterpunch.org/2017/05/26/ukraine-and-the-nato-military-alliance/)

- Washington is negative to Russia
- Author is negative to Washington
- Autor attitude to Russia?

Outline

- Corpus of analytical articles RuSentRel annotated with sentiment attitudes

- Experiments on extracting sentiments with machine-learning methods
  - Baselines
  - Features
  - Human performance in the same task

- Analysis of errors and future research

New sentiment-annotated collection RuSentRel

- Russian translations of political articles in portal inosmi.ru
- Articles contains a lot of opinions and attitudes between entities and author’s attitude towards some entities
- Annotation for whole texts
  - Attitudes between named entities
  - Sentiment from author to named entities
  - Only positive or negative sentiments are labeled
  - Currently: labeling of 73 whole texts

Отношения **Финляндии** и **Швеции** можно считать хорошими. Ведь входили же страны в состав одного королевства до 1809 года. Страны объединяет также и то, что они не входят в **НАТО**, но являются партнерами альянса. Кроме того, **Финляндия** и **Швеция** укрепляют двустороннее сотрудничество в области обороны.

Несмотря на все это, в ходе обсуждений в **Култаранте** возникли разногласия. Бывший министр обороны Швеции **Карин Энстрём** возмутилась тем, что президент **Финляндии** встречается в следующем месяце с президентом **России** в **Хельсинки**. По мнению второго шведского участника дискуссий, **Владимиру Путину** не были бы рады в **Швеции**.

Whole Text Labeling

Обама, Асад, neg (Obama, Asad, neg)

США, ИГИЛ, neg (USA, ISIL, neg)

Иран, Асад, pos (Iran, Asad, pos)

США, Ирак, neg (USA, IRAK, neg)

США, Афганистан, neg (USA, Afganistan, neg)

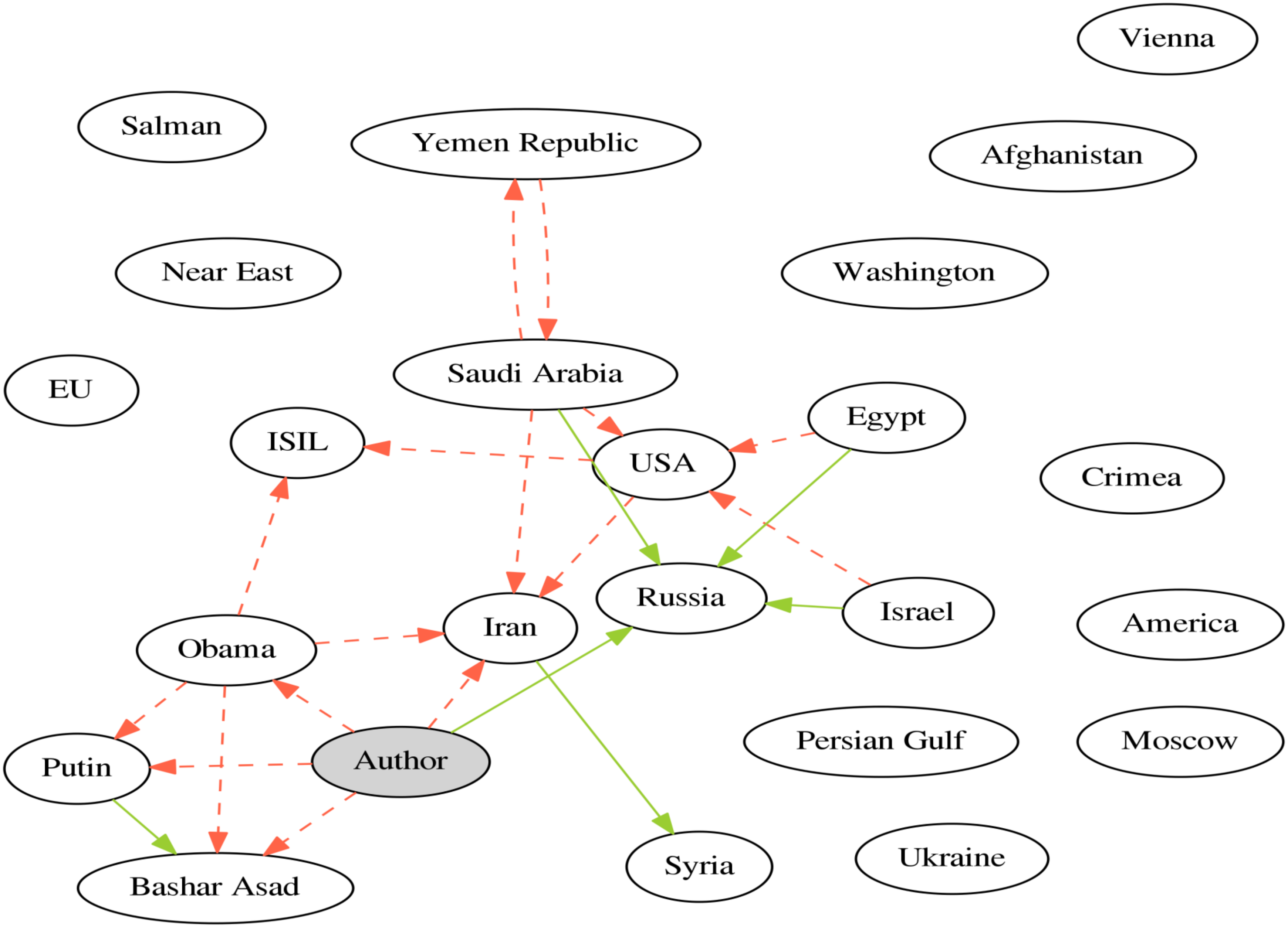
Япония, США, pos (Japan, USA, pos)

Южная Корея, США, pos (South Korea, USA, pos)

Австралия, США, pos (Australia, USA, pos)

author, Обама, pos (Author, Obama,pos)

Picture of the whole text



Corpus annotation

- Attitudes
  - Very low recall: people label much less relations than really mentioned
  - Two students: initial labeling
  - Pooling of labels
  - Meta-annotator: tries to find additional attitudes

- Additionally
  - Automatically labeled named entities (CRF, Mozharova, Loukachevitch, 2016)
  - Synonyms and variants of named entities
    - Путин, Владимир Путин, Владимир Владимирович Путин
    - Биби Нетаньяху, Нетаньяху

Task

- Classification of attitudes between named entities into three classes: positive, negative, neutral
- Measure: averaged sum of F-measure of positive class and negative class

- The first attempt
  - Summer School “Natural Language Processing and Data mining” (2017)
    - Higher School of Economy

Corpus statistics

	Training collection	Test collection
Number of documents	44	29
Avg. number of sentences per document	74.5	137
Avg. number of mentioned entities per doc.	194	300
Avg. number of unique named entities per doc.	33.3	59.9
Avg. number of positive sentiment pairs of named entities per doc.	6.23	14.7
Avg. number of negative sentiment pairs of named entities per doc.	9.33	15.6
Avg. number of neutral sentiment pairs of named entities per doc.	120	276

Use of machine learning

- Only entities mentioned in the same sentences are considered
  - Generation of neutral examples: all pairs of entities mentioned in the same sentences, between which the attitude is not annotated
- Conventional machine learning approaches (sci-kit learn)
  - Naïve bayes, svm, random forest
  - With grid-parameter tuning
- Two types of features
  - Entities features
  - Context features
  - Altogether, 54 features

Entities' features

- word2vec similarity between entities
  - vectors of multiword expressions are calculated as the averaged sum of the component vectors;
- the named entity type according to NER recognizer:
  - person, organization, location, or geopolitical entity;
- the presence in the lists of countries or their capitals;
- the relative frequency of a named entity or the whole synonym group in the document;
- the order of two named entities;
- <u>Concrete lemmas of named entities are not used</u>

Context Features

- the number of sentiment words from RuSentiLex vocabulary:
  - the number of positive words, number of negative words;
  - the average sentiment score of the sentence;
  - the average sentiment score before the first named entity, between named entities, and after the second named entities according to RuSentiLex;
- the distance between named entities in lemmas;
- the number of other named entities between the target pair;
- number of commas between the named entities
- maximum, minimum and average for all features.s

Baselines

Baseline method	Precision	Recall	F-measure
Baseline_neg	0.027	0.39	0.05
Baseline_pos	0.021	0.40	0.04
Baseline_random	0.039	0.215	0.065
Baseline_distr	0.045	0.23	0.075
Baseline_school	0.13	0.103	0.12

Machine learning results

Method	**Precision**	**Recall**	**F1**
KNN	0.18	0.06	0.09
Naïve Bayes Gauss	0.06	0.15	0.11
Naïve Bayes Bernoulli	0.13	0.21	0.16
SVM: Default values	0.35	0.15	015
SVM: Grid search	0.09	**0.36**	015
Random forest **Default values**	**0.44**	0.19	**0.27**
Random forest **Grid search**	0.41	0.21	**0.27**

Machine learning results

Method	Precision	Recall	F1
KNN	0.18	0.06	0.09
Naïve Bayes Gauss	0.06	0.15	0.11
Naïve Bayes Bernoulli	0.13	0.21	0.16
SVM: Default values	0.35	0.15	015
SVM: Grid search	0.09	**0.36**	015
Random forest **Default values**	**0.44**	0.19	**0.27**
Random forest **Grid search**	0.41	0.21	**0.27**
Meta-annotator	0.62	0.49	**0.55**

Analysis of Errors

- ***Лиухто*** *говорит, что он начал склоняться к вступлению Финляндии в* ***НАТО***
- ***Путин*** *хочет войти в историю как царь, расширивший территорию* ***России***
- *Глава комиссии по иностранным делам эстонского парламента Рийгикогу, бывший министр иностранных дел* ***Свен Миксер*** *(Sven Mikser) считает, что, возможно, президент* ***Владимир Путин*** *не стремится присоединить к России в первую очередь страны Балтии, но подобные намерения вполне могут существовать*

Corpus is published

- https://github.com/nicolay-r/RuSentRel/tree/v1.0
- https://github.com/nicolay-r/sentiment-relation-classifiers/tree/dialog_2018

Conclusion

- RuSentRel corpus of annotated analytical articles
  - Sentiment of author to named entities
  - Attitudes of named entities to each other
  - Extracted named entities
  - Synonymic rows of named entities
- Applying “conventional” machine learning methods
  - Relatively low results
  - But baselines are also very low
- Future research
  - Neural networks
  - Automatic generation of training collection