Distant Supervision for Sentiment Attitude Extraction

Nicolay Rusnachenko[1] Natalia Loukachevitch[1,2] Elena Tutubalina[3]

[1]Bauman Moscow State Technical University, Moscow, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
[3]Kazan Federal University, Kazan, Russia

Introduction

News articles often convey attitudes between the mentioned subjects, which is essential for understanding the described situation.
We describe a new approach to distant supervision for extracting sentiment attitudes between named entities mentioned in texts.

- **Example:** "... [USA] is considering the possibility of new sanctions against [Russia] ... ";
- Context illustrates a negative USA→Russia attitude.

How to create a training set?

- Two factors are used for automatic labeling of the news collection (RuAttitudes[1]):
  1. Pair-Based;
  2. Frame-Based.

RuSentiFrames

The **RuSentiFrames-v1.0**[3] lexicon describes sentiments and connotations conveyed with a predicate in a verbal or nominal form.
The structure of the frames includes:

1. Role Designation:
   - A0 is an argument exhibiting features of a Prototypical Agent, and A1 is a Theme (as in PropBank).
2. Dimentions:
   - the attitude of the author of the text towards mentioned participants;
   - positive or negative sentiment between participants;
   - positive or negative effects to participants;
   - positive or negative mental states of participants related to the described situation.

- Assertions is a score of confidence:
  - 1 – is true almost always;
  - 0.7 – considered in default.

Example: Frame "Одобрить" (Approve)

```
"roles":  {"a0":  "who approves",
           "a1":  "what is approved"}
"polarity":  {["a0", "a1", "pos", 1.0],
              ["a1", "a0", "pos", 0.7]},
"effect":  {["a1", "pos", 1.0]},
"state":  {["a0", "pos", 1.0],
           ["a1", "pos", 1.0]}
```

Parameter	Number
Verbs	2 794
Nouns	822
Phrases	2 401
Unique Entries	6 036
Total Entries	6 412
A0 → A1 (positive)	2 252
A0 → A1 (negative)	2 802

Table 1: Quantitative characteristics of RuSentiFrames-v1.0 entries.

RuSentRel

The **RuSentRel**[2] corpus consisted of analytical articles from Internet-portal `inosmi.ru` devoted to international relations.
Annotation:

1. The author's relation to mentioned named entities;
2. The relation of subjects expressed as named entities to other named entities.

RuSentRel Statistics

Parameter	Value
Number of documents	73
Total opinion pairs	1 361
Sentences (avg./doc.)	105.75
Opinion pairs (avg./doc.)	18.64
Positive opinion pairs (avg./doc.)	8.71
Negative opinion pairs (avg./doc.)	9.93

Table 2: Attitude statistics of RuSentRel-v1.1 corpus.

Train Collection Development Steps

1. We use two different methods of sentiment attitude annotation, applied to the news title:
   - **Pair-Based** – utilizing the pre-assigned attitudes organized in a list of pairs;
   - **Frame-Based** – utilizing frame entries from the RuSentiFrames lexicon.
2. We intersect the annotations and separate result:
   - With the **same** polarity;
   - With the **different** polarity according to both sources.

Development Stages Statistics

Corpus	doc. level attitudes	texts count	titles and sentences
Pair-Based	60 788	52 377	136 496
Frame-Based	55 566	43 383	104 205
Intersection	22 589	20 885	50 958
Different	7 929	7 435	17 939
Same$_{\text{RuAttitudes}}$	14 660	13 450	33 019

Table 3: The statistics of automated annotation of texts and sentences.

RuAttitudes Development Flow

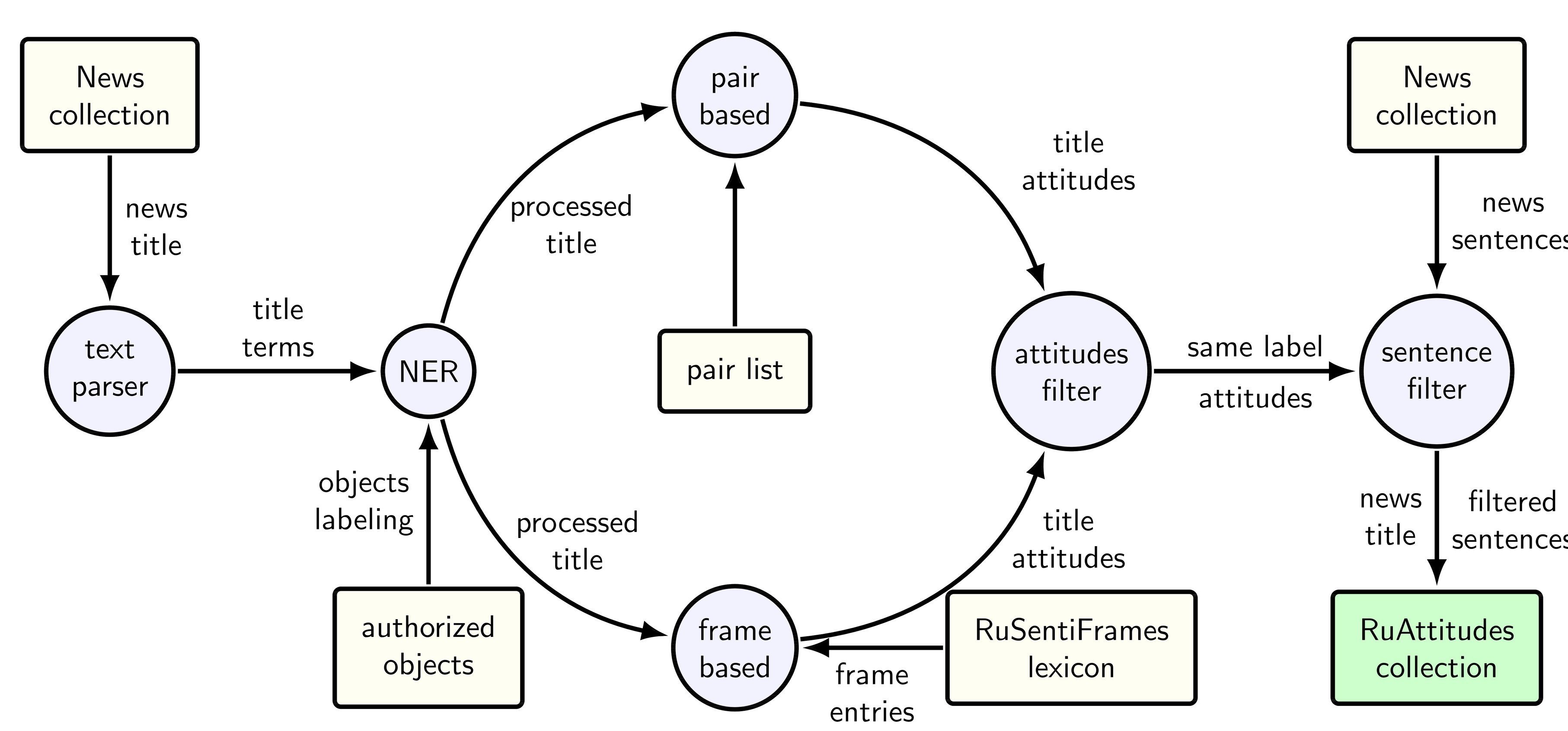

Figure 1: Train collection development flow

Text Processing and Embedding

News processed: 2.8×10^6

1. Text processing involves:
   - Tokenization to demarcate text string into words and punctuation signs;
   - Numbers and URLs masking (considered as non-meaningful).
2. Model input vector includes:
   - Sentence embedding, yields of: words, tokens (set of predefined size);
   - Features, is a randomly initialized vectors:
     - Distance embedding [1];
     - Part-Of-Speech (POS) tags embedding;

Type	Parameters	Values
Tokens	$\langle size, l_t \rangle$	$\langle 17, 10^3 \rangle$
Words	$\langle size, l_w, w \rangle$	$\langle 147 \cdot 10^3, 10^3, 20 \rangle$
POS	v_{size}	5
Distance	v_{size}	5

Table 4: Embedding parameters, where v_{size} is the size of embedding vectors.

Models And Training Types

1. Utilized architectures:
   - CNN – Convolutional Neural Networks;
   - PCNN – Piecewise Convolutional Neural Networks [1].
2. Training types:
   - Single Sentence Training – assumes to predict a sentiment label by a single sentence [1] (CNN, PCNN);
   - Multiple Sentence Training – assumes to predict a sentiment label for sentences set [2] (MI-CNN, MI-PCNN);

Description	Parameters	Values
Minibatch	$\langle n, m \rangle$	$\langle 8, 3 \rangle$
Optimiser	$\langle lr, \rho, \epsilon \rangle$	$\langle 0.1, 0.95, 10^{-6} \rangle$
Terms	k	50
Window size	w	3
Filters count	c	300
Dropout	ρ	0.9

Table 5: Predefined training parameters.

Experiments[4] Description

1. RSR – RuSentRel based dataset with sentence-level attitude labeling;
2. RSR+RA – a combination of RSR and RuAttitudes (RA) datasets.

Parameter	RA	RSR
Documents	13 450	73
Opinions on sentence level	35 125	2 879
– Negative	26 904	1 602
– Positive	8 221	1 277
Avg. opinions per sentence	1.06	**2.26**
Avg. sentences per opinion	2.40	2.57

Table 6: Comparison of RuAttitudes and RuSentRel based (RSR) datasets for experiments.

	RSR			RSR+RA		
Models	F_1	P	R	F_1	P	R
neg$_{\text{baseline}}$	.39	.31	.54	.39	.31	.54
rand$_{\text{baseline}}$	.49	.51	.48	.49	.51	.48
CNN	.52	.52	.55	.63	.62	.66
PCNN	.59	.58	.61	.67	.66	.69
MI-CNN	.57	.56	.60	.62	.60	.65
MI-PCNN	**.62**	.60	.64	**.68**	.67	.70

Table 7: Result of single sentence (CNN, PCNN) and multiple sentence (MI-CNN, MI-PCNN) trained models in following experiments: RSR – RSR results trained on RSR; RSR+RA – RSR results trained on RSR+RA.

Conclusion

This result analysis demonstrates the model classification improvements achieve 13.4% increase in F_1 when the latter being trained with the developed collection.

Related Works

[1] Nicolay Rusnachenko and Natalia Loukachevitch.
Neural network approach for extracting aggregated opinions from analytical articles.
In *International Conference on Data Analytics and Management in Data Intensive Domains*, pages 167–179. Springer, 2018.

[2] Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang.
Relation extraction with multi-instance multi-label convolutional neural networks.
In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480, 2016.

Links

1. `https://github.com/nicolay-r/RuAttitudes/tree/v1.0`
2. `https://github.com/nicolay-r/RuSentRel/tree/v1.1`
3. `https://github.com/nicolay-r/RuSentiFrames/tree/v1.0`
4. `https://github.com/nicolay-r/attitudes-extraction-ds`