Sentiment analysis algorithms for the Belarusian NooJ module in touristic sphere
Co-authors

Yuraś Hiecevič

Aliena Kryvalcevič

Anastasija Drahun

Jaŭhienija Zianoūka

Nastassia Kazloŭskaja

Aliaksandr Ščarbakoŭ
Background of Belarusian NooJ module 2011-2015

- 2011 Overview of Belarusian and Russian grammatical dictionaries
- 2012 First Belarusian, Russian *.nod dictionaries, publication of Belarusian NooJ module with texts, projects, grammars, dictionaries
- 2013 Syntax grammars for QEMU and direct speech processing
- 2014 Next steps in syntax grammars for QEMU processing
- 2015 First morphological and syntax grammars for
  - syllabification,
  - phonemic
  - and prosodic transcriptions.
  - Foundation of Belarusian NooJ team
  - Minsk NooJ conference
Background of Belarusian NooJ module 2016 (first half)

- 2016 Next steps in morphological and syntax grammars for syllabification, phonemic and prosodic transcription;
  - generating of texts for robotics command support;

Robot & Belarusian TTS
Background of Belarusian NooJ module 2016 (second half) – 2017

- **Weekly news letters** for Belarusian NooJ team and Max Silberstein about done tasks and tasks that we are going to do
  - Hanna Stanislavenka
- **Belarusian video tutorials for NooJ**
  - Julia Baradzina, Maryna Marchyk
- **Sentiment analysis** for Belarusian touristic sphere
- **New Java NooJ core** implementation
  - Alaksandar Scharbakou, Yuras Hetsevich
Belarusian video tutorials for NooJ
12 videos in Belarusian for new NooJ participants

• https://www.youtube.com/playlist?list=PLtc_R9i0zr6QjyLk5_Vn_9F4balHK42XW
Sentiment analysis algorithms for the Belarusian NooJ module in touristic sphere
Plan in general

• Overview of the problem and task definition
• Corpora collecting: training and test
• Experiments: 1st, 2nd on basis of training corpora
• Evaluations: on basis of training corpora and final testing corpus
• Future plans
Overview of the sentiment analysis (SA) problem, focusing to Belarus

• Sentiment analysis aims to determine the opinion of a speaker, writer, or other subject with respect to some topic (emoticons “)))))”, good, bad, sad words about object)

• Belarus becomes more welcome and open for touristic flow

• How can we measure opinion of newcomers about Belarus?
There is no any SA software for Belarusian touristic content... (:
We hope that NooJ can help us!

:)
Task definition

• Sphere – tourism
• Main topic of reviews – Mirski castle (Мірски замак)
• Sources – Internet pages and social networks hash tags with reviews for main topic
Sources – more than 20 web links

http://belarusgid.by/mirskij-zamok/
https://forum.onliner.by/viewtopic.php?t=244845&start=0
http://mircki-zamok.relax.by/otzyvy/
https://turzametka.com/mirskij-zamok/
https://tonkosti.ru/otzyv/
https://mirzamak.tam.by/otzyvy/
https://vladkonst.tourister.ru/photoalbum/31112
https://www.mishanita.ru/2014/08/11/23166/
http://citykey.net/review/priehav-v-eto-mesto-popadaesh-v-srednevekove
https://govorim.by/strana-belarus/93107-mirskiy-zamok.html
https://www.tuda-suda.by/countries/europe/belarus/mir/mirskij-zamok
...
Examples of reviews

POSITIVE
• “Замок очень красивый и внутри, и снаружи. Впечатляющие виды днем и вечером с подсветкой. Рядом озеро и небольшой парк с усыпальницей”

NEGATIVE
• “При въезде в город сам замок не виден, указателей никаких нет. Немножко заплутали, но благо город небольшой, поэтому быстро нашли. Конечно замок вживую не производит величественного впечатления. Но при покупке билета нам не было предложено аудиогида. (Обязательно просите, потому что без него или без экскурсовода там нечего делать). Экспозиция очень слабенькая.”
Corpora collecting (txt files)

- 1st training corpus
  - pos. reviews – 503
  - neg. reviews – 168

- 2nd training corpus
  - Additional reviews were added to the 1st training corpus
  - pos. reviews – 634
  - neg. reviews – 207

- Test corpus
  - Reviews are not appear in training corpora
  - pos. reviews – 83
  - neg. reviews – 24
Transferring training and test corpora into NooJ corpora
Experimental grammars under the basis of the 1\textsuperscript{st} training corpus

Naïve variant

More accurate variant
One of the “Positive” subgraph

Positive phrases (true answers) – 348

[Image of a diagram showing positive phrases in Russian]
One of the “Negative” subgraph

Negative phrases (true answers) - 98
Concordance for the 1st experiments
Evaluation of the 1\textsuperscript{st} experiments

- (automatically calculated in the MS Excel)
- Precision: 0.91. Recall: 0.16. F-measure: 0.27.

- **Discussion** – It works! But it is not good accuracy.
- **Decision** – we have to
  - update “true answers” because recall is too low
  - and check it under 2\textsuperscript{nd} corpus for training
Charts for the 1\textsuperscript{st} experiment data analysis

**Amount of “true answers”**

1 evaluation

<table>
<thead>
<tr>
<th>POS</th>
<th>NEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>100</td>
</tr>
</tbody>
</table>

**Accuracy**

Accuracy, %

<table>
<thead>
<tr>
<th>F-measure</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>85</td>
</tr>
</tbody>
</table>
Experimental grammars on the basis of the 2^{nd} training corpus
True answers for subgraphs of the 2nd training corpus

- Positive phrases – 524
- Negative phrases – 404
Evaluation of the 2\textsuperscript{nd} experiment

- **Precision:** 0,92. **Recall:** 0,89. **F-measure:** 0,90

- **Discussion** – It is good accuracy! Can we use it?

- **Decision** – try to test grammar from 2\textsuperscript{nd} experiment under the test corpus
Charts for the 1\textsuperscript{st} and the 2\textsuperscript{nd} experiments data analysis

Improving of “right answers” subgraphs

<table>
<thead>
<tr>
<th></th>
<th>1 test</th>
<th>2 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>98</td>
<td>348</td>
</tr>
<tr>
<td>NEG</td>
<td>0</td>
<td>404</td>
</tr>
</tbody>
</table>

Accuracy

Accuracy, %

- F-measure
- Recall
- Precision
Evaluation of 2\textsuperscript{nd} grammar under the test corpus

- Precision: 0.94. Recall: 0.70. F-measure: 0.80

- Discussion – recall is going to down, but precision – to up. F-measure – is going to down but it is still enough.

- Strategic decision – to do
  - visualization of all data in a charts
  - analytical plans for the stabilization of accuracy

<table>
<thead>
<tr>
<th>ДАДЖЕНЬЗ Э КАНКАРАНСА НУДЖ</th>
<th>ВЫДАЛАЕМ ПАУТОРЫ</th>
<th>ВЫВОДЩИК</th>
<th>КАЛЯСКИСЬ</th>
<th>СПАРАЙДИС</th>
<th>СПАРАЙДИС</th>
<th>79</th>
<th>68</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>neg 1 not</td>
<td>как-то</td>
<td>не очень понравилось</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>neg 1 not</td>
<td>кафешки и</td>
<td>недорогой</td>
<td>POS</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>neg 15 not</td>
<td>Ольга крутая</td>
<td>красивая, пока не бросила</td>
<td>neg</td>
<td>neg1</td>
<td>neg1</td>
<td>neg</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>neg 18 not</td>
<td>Минус</td>
<td>Мировой замок</td>
<td>NEG</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>neg 18 not</td>
<td>до него</td>
<td>не очень удобно</td>
<td>NEG</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>neg 19 not</td>
<td>чувствуется ста</td>
<td>GSPLIT</td>
<td>POS</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Charts for the 1\textsuperscript{st}, 2\textsuperscript{nd} and test experiments data analysis

### Improving of “right answers” in subgraphs

<table>
<thead>
<tr>
<th></th>
<th>1 test</th>
<th>2 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>348</td>
<td>524</td>
</tr>
<tr>
<td>NEG</td>
<td>98</td>
<td>404</td>
</tr>
</tbody>
</table>

### Improving of accuracies between experiments

<table>
<thead>
<tr>
<th></th>
<th>1st, %</th>
<th>2nd, %</th>
<th>test, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy (1st)</td>
<td>22</td>
<td>48</td>
<td>98</td>
</tr>
<tr>
<td>Accuracy (2nd)</td>
<td>48</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>Accuracy (test)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Plans for SA grammars improving & using in applications

- to add `<WF>` and `<P>` modifications into grammars for more accurate work grammars in general case
- to improve the constructed syntax grammars algorithm with adding the to a subgraphs of tagged true answers for Adverbs and Adjectives
- to do modifications in Excel calculations when we obtains multiple POS (NEG) annotations for the same review
- to add a tag for neutral state
- to build online Java Belarusian NooJ prototype for sentiment analysis tourism task automatization
Implementation of new Java NooJ core
Statement of the problem

• to develop new dictionary module for NOOJ
• to develop new RegExp module for NOOJ
• to test it under the test set that includes short text, 6 tasks for NooJ Linguistic Analysis & Locate pattern functionality of NooJ formalism
What we can use?

- C# NOOJ version as working program
- Open source of [Java NOOJ](https://www.meta-share.org/) from META-SHARE for modification
- NooJ manual as hand book of NooJ formalizm
- Sample files of *.dic, *.nof, _properties from the Belarusian NooJ module
- Model-View-Controller approche
- E-mails to Max :)


Data model classes diagram
View

• Simple console output
Controller and data models classes diagram
Descriptions of the developed classes

• **Class AnalisedWord**, are presented in the class diagram implements a full linguistic information stored analyzed word contains fields such as: word form, a particular form of words in the text, some of the language and other linguistic information. Appropriate methods are set and returns the value fields of the class. Just overridden method in the object recording line

• **Class LexicalInformation** shown in the class diagram implements a data model and stores the basic linguistic information word as described in the dictionary. It includes fields such as: word form, part of speech, and other linguistic information. Appropriate methods are set and returns the value fields of the class. Just overridden object recording method in a row.

• **Class Paradigm** shown in the class diagram implements a data model for storage in the computer memory all inflected grammar words. It includes fields such as: word form, the number of changes the sign at the end of words, the variable part of the word. Appropriate methods are set and returns the value fields of the class. Just overridden method in the object recording line
• **Class Property** shown in the class diagram implements a data model and stores the additional linguistic information. It includes fields such as: part of the language and other linguistic information. Appropriate methods are set and returns the value fields of the class. Just overridden object recording method in a row.

• **Class SercherOutputPrinter** shown in the class diagram implements a data model and stores the context in which the word is used in the text. It includes fields such as: context before the word, the word, the context after the word. Appropriate methods are set and returns the value fields of the class. Just overridden method in the object recording line.

• **Class Gramm** presented data model to store in the computer memory all inflected grammar words. It includes fields such as: word form, all forms of words with corresponding linguistic information. Appropriate methods are set and returns the value fields of the class. Just overridden method in the object recording line.

Below The the group of classes that belonging to the data controllers: DefWorker, DicWorker, NofWorker, Sercher, Analisys.
• **TextWorker class** provides methods that implement the file processing function of the text type files.

• **Analysis class** provides an implementation of secondary processing of the text prior to analysis, the construction of grammars, text analysis, the preparation of the analysis results to the output file, the output of the analysis of the text in the file.
For the analysis of the text are used 3 types of files for dictionary

- «Dictionary»
- «Inflectional / Derivational description»
- «Dictionary properties definition»
3 classes have been implemented for processing dictionaries.
Description of the developed classes for work with dictionaries

- **DefWorker class** provides methods that implement the file processing function of the type «Dictionary properties definition».
- **DicWorker class** provides methods that implement the file processing function of the type «Dictionary».
- **NofWorker class** provides methods that implement the file processing function of the type «Inflectional / Derivational description».
Analysis class was implemented for word processing

Analysis class provides an implementation of secondary processing of the text prior to analysis, the construction of grammars, text analysis, the preparation of the analysis results to the output file, the output of the analysis of the text in the file.
One simple example how to work new RegExp

Task: Locate pattern: <VERB+Present>

For text: актрыса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для таго каб ускочыць. актрысы ідучь на ўваходы тэатраў.

Must be

Answer is:

Approve: yes
Task 1. Locate pattern: <NOUN>
Text: актриса сядзела і глядзела на йваход.
актрис з аб'езду ішла, для таго каб ускочыць.

<table>
<thead>
<tr>
<th>Text</th>
<th>Before</th>
<th>Seq.</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>актриса сядзела і глядзела на йваход. актрис з аб'езду ішла, для таго каб ускочыць</td>
<td>сядзела і глядзела на йваход. актрис з аб'езду ішла, для таго каб ускочыць</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Answer is:

Approve: yes. Comments: there are some mistakes in letters case
Task 2. Locate pattern: < VERB >

**Text:** актриса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для таго каб ускочыць.

**Answer must be:**

<table>
<thead>
<tr>
<th>Before</th>
<th>Seq.</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>актриса</td>
<td>сядзела</td>
<td>і глядзела на ўваход. актрис</td>
</tr>
<tr>
<td>актриса сядзела і</td>
<td>глядзела</td>
<td>на ўваход. актрис з аб</td>
</tr>
<tr>
<td>ўваход. актрыс з аб'езду ішла</td>
<td>для таго каб ускочыць.</td>
<td></td>
</tr>
</tbody>
</table>

**Answer is:**

**Approve:** yes. **Comments:** please check small mistakes in punctuation
Task 3. Locate pattern: <актрысамі>

- Text: актрыса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для таго каб ускочыць.
- Answer must be:

<table>
<thead>
<tr>
<th>Text</th>
<th>Before</th>
<th>Seq.</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>актрыса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Answer is:

- Approve: yes. Comments: please check case of letters
Task 4. Locate pattern: <актрысамі><VERB>

- Text: актрыса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для таго каб ускочыць.
- Answer must be:

- Answer is:

- Approve: no. Comments: must be other word located.
Task 5. Locate pattern: <VERB+Present>

For text: актрыса сядзела і глядзела на ўваход. актрыс з аб'езду ішла, для таго каб ускоочыць. актрысы ідуць на ўваходы тэатраў.

Must be

Answer is:

Approve: yes
Task 6. Locate pattern: <актрысы><VERB+Present>

• For text from task 5
• Answer must be:

<table>
<thead>
<tr>
<th>Text</th>
<th>Before</th>
<th>Seq.</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>ишла, для таго каб ускочыць. актрысы идуць на ўваходы тэатраў.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Answer is:

• Approve: yes.
Plans for further development of new core of Java NooJ

• To complete the RegExp module. Re-check tests.
• To finalize the dictionary module. Make tests.
• To extend the functionality for NooJ grammars.
KrokApp application in touristic sphere

Cities

- Viciebsk
- Homieĺ
- Mahilioŭ
- Lida
- Baranavichy
- Orsha
- Babruisk
- Barysaŭ

Church of St. Simon and St. Helena (also known as the Red Church) is Catholic parish church in the Neo-Romanesque style in Minsk.
Yuras Hetsevich
PhD, Head of the Laboratory of Speech Synthesis and Speech Recognition UIIP NAS Belarus

yuras.hetsevich@gmail.com
+375 29 681 52 53
Belarus, Minsk

Thank you for your attention!

Questions? Suggestions? Cooperation?