Coreference Resolution in Russian: State-of-the-Art Approaches Application and Evolvement

Speaker: Andrey Sysoev
Ivan Andrianov
Alexandra Khadzhiiskaia

Moscow, 2017
Let's start with an example

If a **bulb** in your **car** burned out – change **it**.

Если в **машина** перегорела **лампа** – замените **её**.

*What should I actually change?*
Let's start with an example

If a **bulb** in your **car** burned out – change it.

Если в машине перегорела лампа – замените её.

*What should I actually change? A car?*
Let's start with an example

If a bulb in your car burned out – change it.

Если в машине перегорела лампа – замените её.

What should I actually change? A car? Or just a bulb?
Coreference resolution. Where to use?

- Relation extraction
- Question-answering systems
- Sentiment analysis
Coreference resolution in practice

Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.
Coreference resolution in practice

Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.
Coreference resolution in practice

Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.
Coreference resolution in practice

Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.
Coreference resolution in practice

Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.

Mikhail Lomonosov Russian scientist his

One of his discoveries atmosphere of Venus
Antecedent and anaphor

Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.
Antecedent and anaphor

Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.

Михаил Васильевич Ломоносов —— his —— Михаил Васильевич Ломоносов

? —— его —— ?
Antecedent and anaphor

Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.
Evaluation.

Test corpus and metrics

Corpus:

- RuCor from RuEval-2014
- 181 (179 after fixing conflicting markup) documents

Quality measures:

- Precision / Recall / F1
- MUC / B3 / CEAF_{entity} / CEAF_{mention}

10-fold crossvalidation
Preprocessing

One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.

Texterra: https://api.ispras.ru/demo/texterra
Mention detection.
What is mention?

One of his discoveries is the atmosphere of Venus.

Одно из его открытий — атмосфера Венеры.
Mention detection

- Identify mention heads
- Expand heads to full mentions

A Joint Framework for Coreference Resolution and Mention Head Detection
Head identification

One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
Head identification. Features

- Internal morphological
  - POS-tag
  - number
  - gender
  - animacy
- Syntax
  - position within sentence
  - relations of a token
- Syntactic context - morphological features for syntactic parent
- Context - basic morphological features for neighbors
- Frequency - TF weighting
- Semantic
Head identification. Semantic features

- Groundtruth heads from training documents are clustered (word2vec, k-means).
Head identification. Semantic features

- Groundtruth heads from training documents are clustered (word2vec, k-means).
- Features - distance and similarity from head candidate to clusters centroids.
One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
Head identification. Evaluation

![Bar chart showing evaluation metrics for different cluster sizes and semantic feature conditions.](image-url)

- **Y-axis:** Values range from 0.62 to 0.74.
- **X-axis:** Conditions include:
  - Without semantic features
  - With 105 clusters
  - With 110 clusters
  - With 115 clusters
  - With 120 clusters
  - With 125 clusters

- **Legend:**
  - Blue: Precision
  - Red: Recall
  - Yellow: F1-measure

- **Metrics:**
  - Without semantic features: 0.7326 Precision, 0.6628 Recall, 0.7049 F1-measure
  - With 105 clusters: 0.7289 Precision, 0.6839 Recall, 0.7049 F1-measure
  - With 110 clusters: 0.7288 Precision, 0.6841 Recall, 0.7049 F1-measure
  - With 115 clusters: 0.7289 Precision, 0.6842 Recall, 0.7049 F1-measure
  - With 120 clusters: 0.728 Precision, 0.6839 Recall, 0.7046 F1-measure
  - With 125 clusters: 0.7288 Precision, 0.6838 Recall, 0.7049 F1-measure
Mention detection

Identify mention heads

Expand heads to full mentions

A Joint Framework for Coreference Resolution and Mention Head Detection
Head expansion

One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
One of his discoveries is the atmosphere of Venus.

Одно из его открытий — атмосфера Венеры.
One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
Head expansion

One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
Head expansion

One of his discoveries is the atmosphere of Venus.

Одно из его открытий – атмосфера Венеры.
Head expansion. Features

- **Token-based** [head, candidate token, nearest neighbours]
  - word form
  - lemma
  - POS-tag
- **Position-based**
  - direction from head to candidate
  - distance between head and candidate
  - head/candidate is the first/last token of the sentence
- **Context-based**
  - head and candidate are parts of the same named entity
  - head/candidate is a syntactic ancestor of candidate/head
  - POS-tag pattern for words between head and candidate
Head expansion. Evaluation

![Bar chart showing evaluation results for different head expansion methods.](chart.png)
Coreference resolution. Easy-First Mention Pair algorithm

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - his
One of his discoveries - atmosphere of Venus
    Mikhail Lomonosov - his
    Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
    his - atmosphere of Venus

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - his
One of his discoveries - atmosphere of Venus
Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

Mikhail Lomonosov
Russian scientist

his

atmosphere of Venus
One of his discoveries

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - his
atmosphere of Venus

Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
his - atmosphere of Venus

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - atmosphere of Venus

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs (antecedent-anaphor)

Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

Mikhail Lomonosov

Russian scientist

his

atmosphere of Venus

One of his discoveries

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs (antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - atmosphere of Venus

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

Mikhail Lomonosov - atmosphere of Venus
Russian scientist - One of his discoveries

A State-of-the-Art Mention-Pair Model for Coreference Resolution
Coreference resolution.
Mention-pair classifier features

- Basic linguistic: word forms, lemmas, part-of-speech tags, grammemes (gender, number, animacy) for mention head and context words.
- Grammemes agreement: mention heads share the same key grammemes (number, gender, animacy, pro-nominality).
- Positional: distance, place within sentence boundaries.
- Named entity: mention types and their agreement.
- Structural: mention size and interrelation with other mentions of the text.
- Surface form matching: lexicographic similarity and textual representation equality indicators.
- Syntactic: grammar role, sharing same parent node or clause.
Jaccard Item Set mining

A State-of-the-Art Mention-Pair Model for Coreference Resolution

Segond M., Borgelt C. (2011)
Item Set Mining Based on Cover Similarity
Jaccard Item Set mining

<table>
<thead>
<tr>
<th></th>
<th>ANT.PER</th>
<th>ANAF.PER</th>
<th>LEX_SI&gt;M0.5</th>
<th>NUM</th>
<th>ANIM</th>
<th>SENT_ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TRUE</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

A State-of-the-Art Mention-Pair Model for Coreference Resolution

Segond M., Borgelt C. (2011)
Item Set Mining Based on Cover Similarity
Jaccard Item Set mining

<table>
<thead>
<tr>
<th></th>
<th>ANT_PER</th>
<th>ANAF_PER</th>
<th>LEX_SIM&gt;0.5</th>
<th>NUM</th>
<th>ANIM</th>
<th>SENT_ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TRUE</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

Segond M., Borgelt C. (2011)
*Item Set Mining Based on Cover Similarity*
### Jaccard Item Set mining

<table>
<thead>
<tr>
<th></th>
<th>ANT_PER</th>
<th>ANAF_PER</th>
<th>LEX_SIM&gt;0.5</th>
<th>NUM</th>
<th>ANIM</th>
<th>SENT_ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>TRUE</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>FALSE</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TRUE</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

\[\text{ANT\_PER} \& \text{ANAF\_PER} \& \neg(\text{LEX\_SIM}>0.5)\]

---

*Uryupina O., Moschitti A. (2015)*  
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

*Segond M., Borgelt C. (2011)*  
*Item Set Mining Based on Cover Similarity*
Coreference resolution. Selecting classifier

![Bar chart showing performance of different classifiers in coreference resolution.](chart.png)
Coreference resolution. Ablation analysis
Coreference resolution.
In the wild

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>0.4768</td>
<td>0.3741</td>
<td>0.4189</td>
</tr>
<tr>
<td>B3</td>
<td>0.4104</td>
<td>0.2957</td>
<td>0.3431</td>
</tr>
<tr>
<td>CEAF_mention</td>
<td>0.4024</td>
<td>0.3702</td>
<td>0.3854</td>
</tr>
<tr>
<td>CEAF_entity</td>
<td>0.2525</td>
<td>0.3433</td>
<td>0.2906</td>
</tr>
</tbody>
</table>
Future work

- Experiments with more machine learning algorithms and approaches.
- Using various clustering algorithms for word embeddings.
- Detailed analysis of features, assumed useless in ablation experiments.
- Tuning coreference resolution algorithm for different mention types.
Credits

Alexandra Khadzhiiskaia

Ivan Andrianov
Thank you!