Coreference Resolution for Russian: Taking Stock and Moving Forward

Alexandra Khadzhiiskaia,
Andrey Sysoev

Moscow 2017
Introduction

• Coreference intuitive explanation:
  Identifying all real world entities mentioned throughout the text.
Example

Шариков злобно покосился на профессора, а он отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Филлип Филлипович остался один в своем кабинете. Он начал мерять комнату.

Sharikov gave the professor an angry look, and he returned him a sideways glance. Ten minutes later Sharikov left for the circus. Philip Philipovich was alone in his cabinet. He started pacing the room.
Шариков злобно покосился на Филлипа Филлиповича, а он отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Профессор остался один в своем кабинете. Он начал мерять комнату.

Sharikov gave Philip Philipovich an angry look, and he returned him a sideways glance. Ten minutes later Sharikov left for the circus. The professor was alone in the cabinet. He started pacing the room.
Terminology

- **Mention** — several words from text that denote an entity
- **Antecedent** — a mention with already established referent
- **Anaphor** — a mention referring to an earlier occurring antecedent
Шариков злобно покосился на Филлипа Филлиповича, а он отправил ему косой взгляд. Через десять минут Шариков уехал в цирк. Профессор остался один в кабинете. Он начал мерять комнату.

• Шариков ← ему
• Sharikov ← him
• Филлипа Филлиповича ← Профессор
• Philip Philipovich ← The professor
Brief history

• A well researched area for English:
  • Methods vary from manually compiled rule-based structures to machine learning algorithms

• Machine learning methods evolved from the most basic to complex

• A great variety of clustering techniques including partitions on whole text
Coreference for Russian

• A shared task on coreference resolution for Russian in 2014 as a part of Dialog Evaluation

• Following papers:

  • Toldova & Ionov 2017: “Coreference resolution for Russian: the impact of semantic features”

  • Sysoev & Andrianov & Khadzhiiskaia 2017: “Coreference resolution in russian: State-of-the-art approaches application and evolvement”
Data and metrics

• RuCor — a corpus of texts from various genres compiled in 2014

• Corpus statistics
  • 179 texts
  • 3,354 chains
  • 15,764 mentions

• Metrics: versions of Precision/Recall/F1
  • MUC
  • B3
  • CEAF_{entity}
  • CEAF_{mention}
Baseline

- Two step process from our previous work*
  - Mention pair classification
  - Clustering
- Adaptations made:
  - Different scheme for syntactic preprocessing
  - Classifier tuning

Baseline: classification

- Random Forest Classifier
- Trained on antecedent-anaphor pairs from RuCor
- Negative examples for training: every anaphor with every mention between itself and its antecedent
- Testing scenario pair generation: a pre-set window of preceding mentions
- Saving all pairs with classifier confidence
Baseline: clustering

• Easy-First Mention Pair algorithm*

Sharikov₁ gave Philip Philipovich an angry look, and he returned him a sideways glance. Ten minutes later Sharikov₂ left for the circus. The professor was alone in the cabinet. He started pacing the room.

• Sharikov₁ — Sharikov₂
• Sharikov₁ — him
• Sharikov₁ — Philip Philipovich
• Philip Philipovich — he
• him — Sharikov₂
• he — him
• Philip Philipovich — The professor

• {Sharikov₁, him, Sharikov₂}
• {Philip Philipovich, The professor, he}

* O. Uryupina and A. Moschitti, “A state-of-the-art mention-pair model for coreference resolution.”
Feature Engineering

- Different types of anaphors:
  - Same lexemes: Sharikov — Sharikov
  - Synonyms: cabinet — room
  - Contextual synonyms: Philip Philipovich — professor
  - Pronouns: Sharikov — him, Philip Philipovich — he
- Pronouns form a special class
Feature Engineering

• Pronouns do not hold any lexical meaning of their own

• Pronouns serve as a referencing mechanism

• Pronouns have shorter referencing scope: about 3 sentences

• Pronoun resolution relies heavily on grammar and distance
Feature Engineering: surface form matching

• Acronym matching: домком — домовый комитет

• Comparison of lemmas representing each mention: him — He —> he

• Different lexicographic similarity measure (strings overlapping, minimum edit distance measure, etc.): professor Philip Preobrazhensky — Philip Philipovich Preobrazhensky
Feature Engineering: surface form matching

• Our suggestion: to filter out these features for mention pairs with one or both pronominal mentions

• Error analysis examples:
  
  • Sharikov gave Philip Philipovich an angry look, and he returned him a sideways glance. The professor was alone in the cabinet. He started pacing the room.

• RFC fails to divide data into groups:
  
  • Pronouns make up a third of all mentions (5078 out of ~15000)
  
  • Misleading features for pronominal group
Feature Engineering: context analysis

• Error analysis examples:
  • Ten minutes later *Sharikov* left for the *circus*. *Philip Philipovich* was alone in *the cabinet*. He started pacing *the room*.

• Adding more features for pronoun resolution

• General idea: is there a better candidate?
  • Analysis of all mentions between currently analysed antecedent and pronoun
Feature Engineering: context analysis

- Boolean matchers for grammatical role, morphological properties, named entities combinations
- Counters for different combinations of same attributes
- Distribution of mentions per sentence in context
Feature Engineering: context analysis

• Ten minutes later Sharikov left for the circus. Philip Philipovich was alone in the cabinet. He started pacing the room.

• Philip Philipovich — Subject + animated + masculine + single + NE:Person

• the cabinet — Indirect Object + inanimated + masculine + single
Feature Engineering: Semantics

- Incorporating semantic information:
  - Semantic similarity between mention head words*

    The professor was alone in the cabinet. He started pacing the room.

- Experiments with filtering for pronominal mention pairs

* S. Toldova and M. Ionov, “Coreference resolution for Russian: the impact of semantic features”
Feature Engineering: Semantics

- Two models tested: word2vec VS fasttext

- Both trained on
  - Russian Wikipedia
  - FactRuEval-2016 corpus
  - LibRuSec sample
  - Blog posts collection

- Dimensionality for both: 100 features vector

- Word2vec trained for lemmas and tokens

- Fasttext trained for tokens
Results

Pronominal experiments

Baseline | Baseline + surface features filtering | Baseline + surface features filtering + context

MUC F1 | B3 F1 | CEAFmention F1 | CEAFentity F1
Results

Semantic similarity experiments

- Baseline + surface features filtering + context
- fasttext
- lemma word2vec
- token word2vec
- fasttext for nouns
- lemma word2vec for nouns
- token word2vec for nouns

MUC F1, B3 F1, CEAFmention F1, CEAFentity F1
Clustering

• One option: previously described EFMP

• More straightforward approach:
  • Take only true pairs
  • Trim them by confidence threshold
  • Unroll into clusters
Clustering: all positive

- Sharikov$_1$ — Sharikov$_2$
- Sharikov$_1$ — him
- Philip Philipovich — he

Confidence threshold

- him — Sharikov$_2$
- Sharikov$_1$ — Philip Philipovich

- \{Sharikov$_1$, him, Sharikov$_2$\}, \{Philip Philipovich, he\}
Clustering: by anaphor

• Combining all antecedents for an anaphor

• Two options:
  • Choose the most confident antecedent
  • Choose the closest antecedent classified as true
Clustering: by anaphor

Sharikov$^1$ gave Philip Philipovich an angry look, and he returned him a sideways glance. Ten minutes later Sharikov$^2$ left for the circus. The professor was alone in the cabinet. He started pacing the room.

- [Sharikov$^1$, Philip Philipovich, he] — him
- [Sharikov$^2$, The professor, the cabinet] — He
Clustering: by anaphor

• Choose the most confident antecedent

  • \([\text{Sharikov}_1, \text{Philip Philipovich}, \text{he}] \rightarrow \text{him}\)

  • \([\text{Sharikov}_2, \text{The professor}, \text{the cabinet}] \rightarrow \text{He}\)
Clustering: by anaphor

• Choose the most confident antecedent
  • \([\text{Sharikov}_1, \text{Philip Philipovich, he}]\) — him
  • \([\text{Sharikov}_2, \text{The professor, the cabinet}]\) — He

• Choose the closest antecedent classified as true
  • \([\text{Sharikov}_1, \text{Philip Philipovich, he}]\) — him
  • \([\text{Sharikov}_2, \text{The professor, the cabinet}]\) — He
Clustering: by anaphor

• Choose the most confident antecedent
  
  • \([\text{Sharikov}_1, \text{Philip Philipovich}, \text{he}]\) — him
  
  • \([\text{Sharikov}_2, \text{The professor}, \text{the cabinet}]\) — He

• Choose the closest antecedent classified as true
  
  • \([\text{Sharikov}_1, \text{Philip Philipovich}, \text{he}]\) — him
  
  • \([\text{Sharikov}_2, \text{The professor}, \text{the cabinet}]\) — He
Clusters: Markov clustering

• Basic idea: to represent classified pairs as a weighted graph.

• Apply Markov clustering algorithm*

• Formula for confidence to weight converting:

\[ w(pair) = \begin{cases} 
2 \times \text{confidence} - 1, & \text{pair is coreferent} \\
0, & \text{otherwise} 
\end{cases} \]

Clusters: Markov clustering

• MCL is a mathematical representation of efficient random walks

• Alternation of two operations:
  • **Expansion** — emulates random walks from each starting point
  • **Inflation** — to establish the boundaries promote already more probable steps from each starting point and demote less probable.

• Final step: unrolling graph into clusters
Clusters: results
Comparison

Future work

• Direct speech boundaries:

Борменталь многозначительно кивнул головой.
- Я тяжко раненный при операции, - хмуро подывывал Шариков, - меня вишение как он отделал, - и он указал голову.

Bormental nodded significantly.
"I was severely wounded in the course of the operation," whined Sharikov. "Look what he did to me," and he pointed to his head. "Are you an anarchist-individualist?" asked Shvonder, raising his brows.
Future work

• Coherent text structure:

В погоне за вожделенным миллионом Бендер не задумывается над тем, что, став обладателем миллиона, он разделит участь Корейко. Бендер с невероятным упорством стремится к обладанию миллионом,

в то время как перед читателем уже полностью прошла судьба Корейко, человека с сорока рублями жалованья и с десятью миллионами в потрепанном чемодане, который он сдает в камеры хранения то одного, то другого вокзала.

In pursuit of the coveted million Bender does not think that, having become the owner of a million, he will share the fate of Koreiko. Bender with incredible tenacity aspires to own a million,

while the reader has already witnessed the fate of Koreiko, a man with forty rubles of salary and with ten millions in a worn suitcase, which he hands over to the storage rooms of station after station.