

Machine-aided multi-document summarization of scientific papers

Student: *Andrey Vlasov*

Research Advisor: *Maxim Panov*

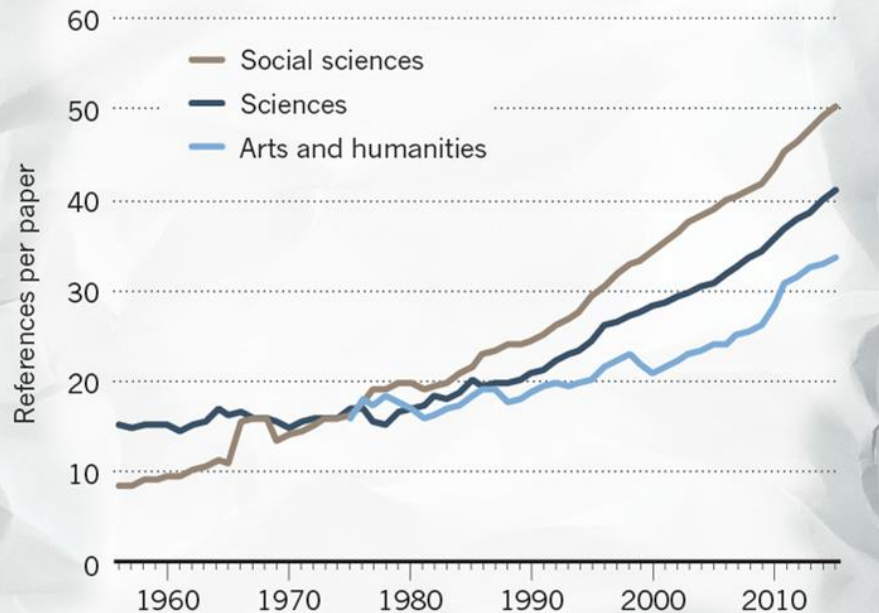
Co-Advisor : *Konstantin Vorontsov*

May, 2020

Stating the problem

References on the rise

The number of references in papers has steadily risen over time, with papers in the sciences now including more than 40 on average.



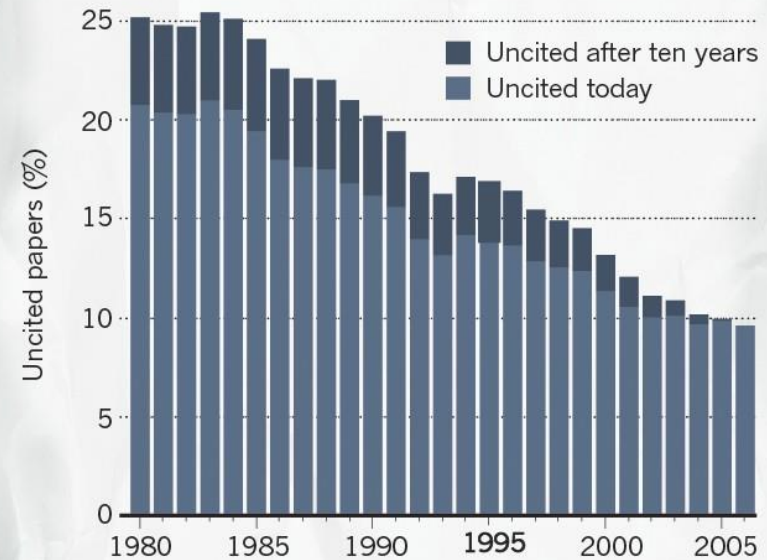
©nature

UNCITED SCIENCE

Data from the Web of Science give an incomplete picture of how much science is never cited: many papers it records as having no citations have actually been cited somewhere.

Downward trend

The share of scientific articles recorded as 'uncited' in each year is falling.



Skoltech

○ **A Deep Reinforced Model for Abstractive Summarization**
Romain Paulus, Caiming Xiong, Richard Socher

Save

Skoltech

○ **A Deep Reinforced Model for Abstractive Summarization**
Romain Paulus, Caiming Xiong, Richard Socher

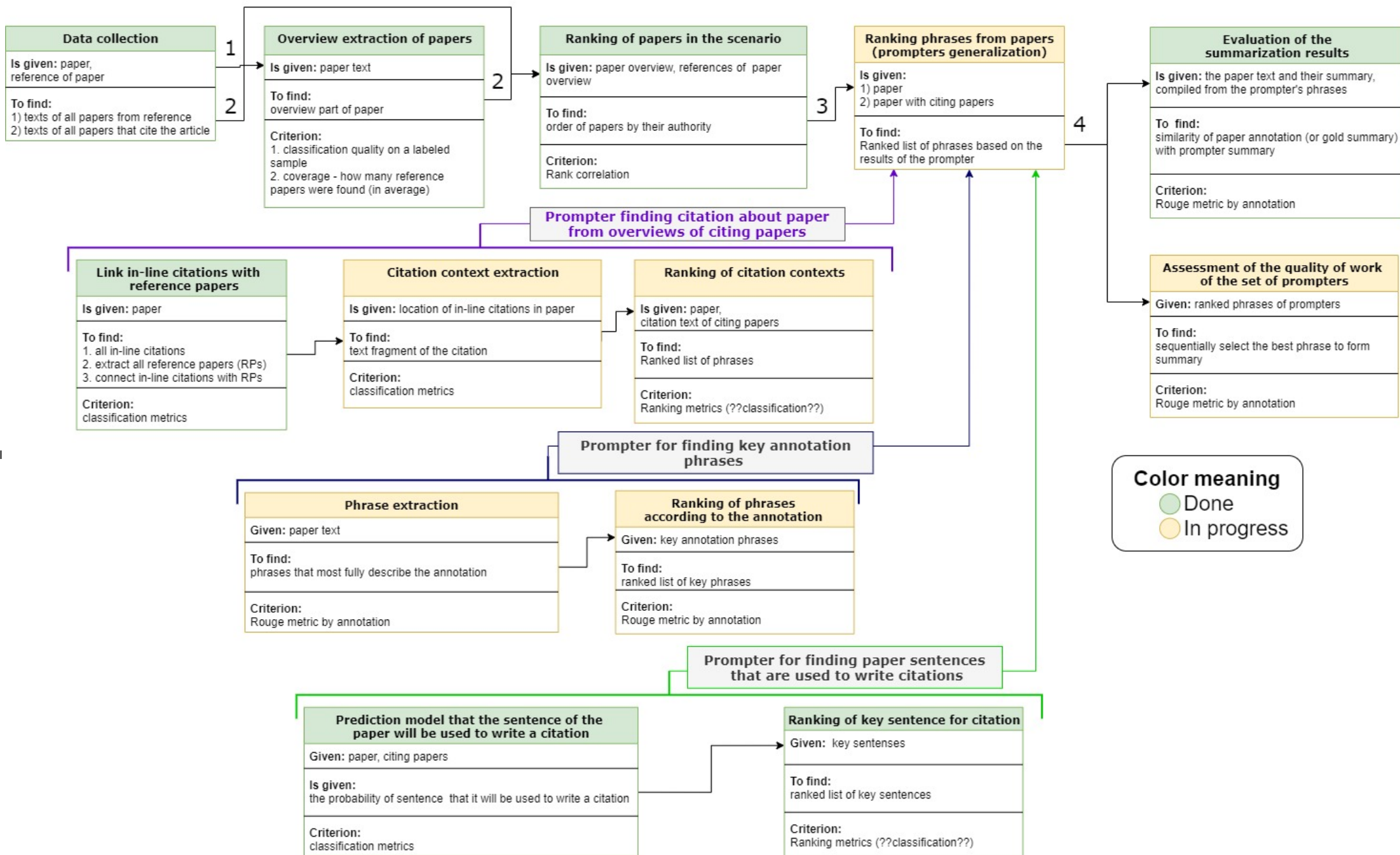
Save

Aim

To create and implement a methodology for solving task of automated multi-document summarization of scientific papers

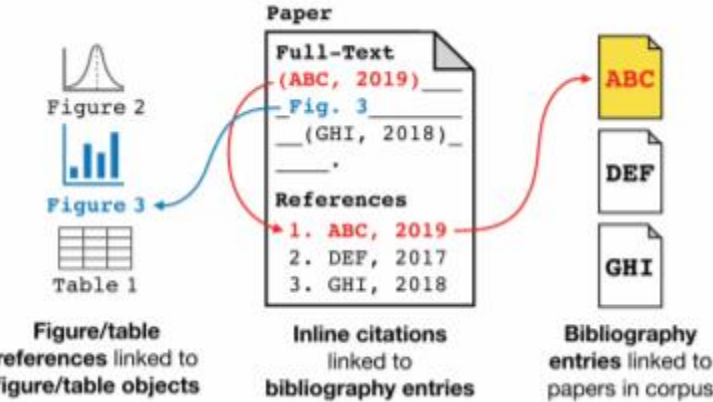
Objectives

1. Data collection
2. Ranking of papers in the collection
3. Getting phrases from prompts (summarization methods) and their ranking
4. Evaluation of the summary & quality of the set of prompts (summarization methods)



Data collection

S2ORC: The Semantic Scholar Open Research Corpus



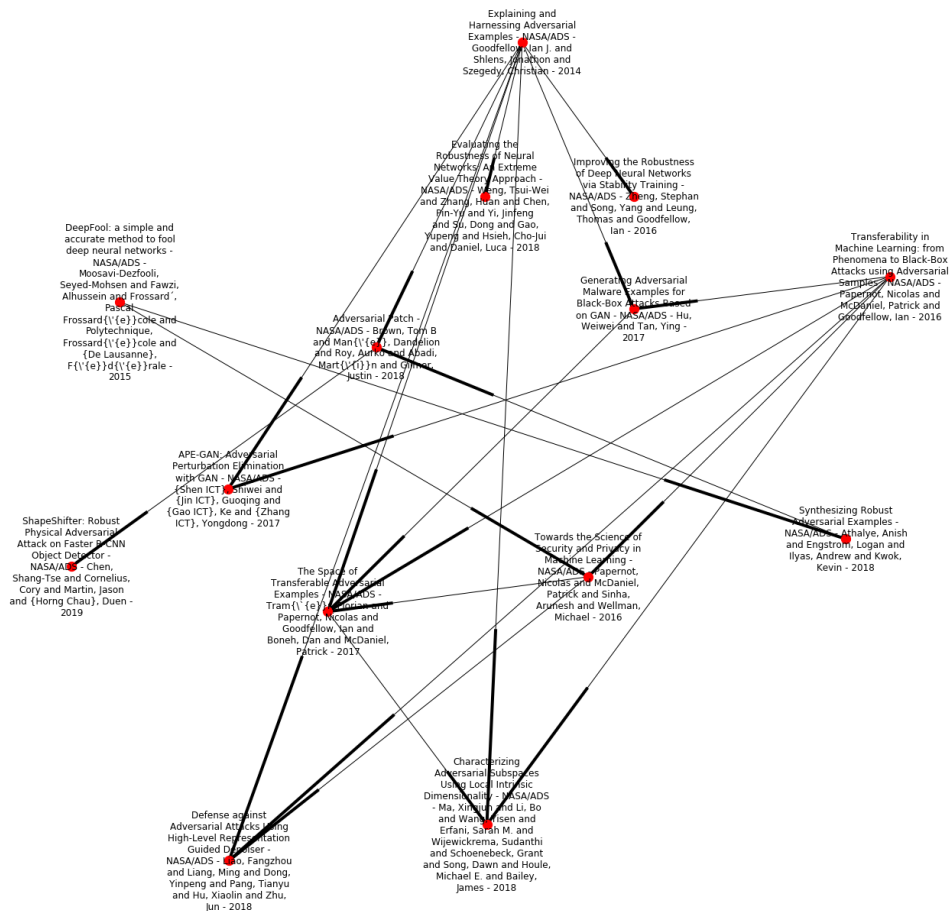
Total papers	81.1M
Papers w/ PDF	28.9M (35.6%)
Papers w/ bibliographies	27.6M (34.1%)
Papers w/ GROBID full text	8.1M (10.0%)
Papers w/ LaTeX full text	1.5M (1.8%)
Papers w/ publisher abstract	73.4M (90.4%)
Papers w/ DOIs	52.2M (64.3%)
Papers w/ Pubmed IDs	21.5M (26.5%)
Papers w/ PMC IDs	4.7M (5.8%)
Papers w/ ArXiv IDs	1.7M (2.0%)
Papers w/ ACL IDs	42k (0.1%)

Table 1. Statistics of papers in this dataset

Data collection

S2ORC: The Semantic Scholar Open Research Corpus

Data collection
Is given: paper, reference of paper
To find: 1) texts of all papers from reference 2) texts of all papers that cite the article



Reference span is a sentence in Reference paper which is mostly cited

Citing paper

Japanese Named Entity Recognition Using Structural Natural Language Processing

Ryohei Sasano*
Graduate School of Information Science
and Technology, University of Tokyo
ryohei@nlp.kuee.kyoto-u.ac.jp

Sadao Kurohashi
Graduate School of Informatics,
Kyoto University
kuro@i.kyoto-u.ac.jp

Abstract

This paper presents an approach that uses structural information for Japanese named entity recognition (NER). Our NER system is based on Support Vector Machine (SVM), and utilizes four types of structural information: cache features, coreference relations, syntactic features and caseframe features, which are obtained from structural analyses. We evaluated our approach on CRI. NE data and obtained a higher F-measure than existing approaches that do not use structural information. We also conducted experiments on IRIEX NE data and an NE-annotated web corpus, and confirmed that structural information

On the other hand, as for English or Chinese, various NER systems have explored global information and reported their effectiveness. In (Malouf, 2002; Chieu and Ng, 2002), information about features assigned to other instances of the same token is utilized. (Ji and Grishman, 2005) uses the information obtained from coreference analysis for NER. (Mohit and Hwa, 2005) uses syntactic features in building a semi-supervised NE tagger.

In this paper, we present a Japanese NER system that uses global information obtained from several structural analyses. To be more specific, our system is based on SVM, recognizes NEs after syntactic, case and coreference analyses and uses information obtained from these analyses and the NER results

On the other hand, various NER systems have explored global information and reported their effectiveness. In (Malouf, 2002; Chieu and Ng, 2002), information about features assigned to other instances of the same token is utilized. (Ji and Grishman, 2005) uses the information obtained from coreference analysis for NER. (Mohit and Hwa, 2005) uses syntactic features in building a semi-supervised NE tagger.

Citance

Citation Marker: Chieu and Ng, 2002

Citation Text: In (Chieu and Ng, 2002), information about features assigned to other instances of the same token is utilized

HasReference:

Reference paper

Lexicon Feature: The string of the token w is used as a feature. This group contains a large number of features (one for each token string present in the training data). At most one feature in this group will be set to 1. If w is seen infrequently during

Suffixes and Prefixes: This group contains only two features: *Corporate-Suffix-List* (for corporate suffixes) and *Person-Prefix-List* (for person prefixes).

4.2 Global Features

Context from the whole document can be important in classifying a name already mentioned previously. A **Reference span** may appear in abbreviated form when it is mentioned again later. Previous work deals with this problem by correcting inconsistencies between the named entity classes assigned to different occurrences of the same entity (Borthwick, 1999; Mikheev et al., 1998). We often encounter sentences that are highly ambiguous in themselves, without some prior knowledge of the

sequences: if two consecutive tokens are matched against bigrams. A list of words occurring more than 10 times in the training data is also collected (*commonWords*). Only tokens with *initCaps* not found in *commonWords* are tested against each list in Table 2. If they are found in a list, then a feature for that list will be set to 1. For example, if *Barry* is not in *commonWords* and is found in the list of person first names, then the feature *PersonFirstName* will be set to 1. Similarly, the tokens w_{t-1} and w_t are tested against each list, and if found, a corresponding feature will be set to 1. For example, if w_{t-1} is found in the list of person first names, the feature *PersonFirstName*_{prev} will be set to 1.

person prefixes like *Mr.*, *Dy.*, etc are not part of person names, whereas corporate suffixes like *Corp.*, *Inc.*, etc are part of corporate names.

4.2 Global Features

Context from the whole document can be important in classifying a named entity. A name already mentioned previously in a document may appear in abbreviated form when it is mentioned again later. Previous work deals with this problem by correcting inconsistencies between the named entity classes assigned to different occurrences of the same entity (Borthwick, 1999; Mihšev et al., 1998). We often encounter sentences that are highly ambiguous in themselves, without some prior knowledge of the entities concerned. For example:

McCann initiated a new global system. (1)
CEO of McCann. (2)

Reference

Reference Span #1: Previous work deals with this problem by correcting inconsistencies between the named entity classes assigned to different occurrences of the same entity
Discourse Facet: methods citation

Dataset

CL-SciSumm

Citance Number: 11 | *Reference Article:* C00-2123.xml | *Citing Article:* J04-4002.xml | *Citation Marker Offset:* ['282'] | *Citation Marker:* Tillmann and Ney 2000 | *Citation Offset:* ['282'] | *Citation Text:* <S sid ="282" ssid = "48">We call this selection of highly probable words observation pruning (Tillmann and Ney 2000).</S> | *Reference Offset:* ['179'] | *Reference Text:* <S sid ="179" ssid = "39">For our demonstration system, we typically use the pruning threshold $t_0 = 5:0$ to speed up the search by a factor 5 while allowing for a small degradation in translation accuracy.</S> | *Discourse Facet:* Method_Citation | *Annotator:* Swastika Bhattacharya |

Example of 1 annotation with 1 citance of reference paper

The extraction of an overview part

Overview extraction of papers
Is given: paper text
To find: overview part of paper
Criterion: 1. classification quality on a labeled sample 2. coverage - how many reference papers were found (in average)

Rule-based / ML approach on features:

- Citation density
- The number of consecutive sentences which include at least 1 citation
- Positional features
 - The section position in the paper
 - An average position of in-line citations in each section

Results:

1. Accuracy = 82% [Gradient Boosting model]
Accuracy = 61% [Rule-based model]
2. Coverage = 57% (percentage of papers included in Overview section which have full-text)

Ranking of papers from the collection

Ranking of papers in the scenario
Is given: paper overview, references of paper overview
To find: order of papers by their authority
Criterion: Rank correlation

Features:

- 1) Year of publication
- 2) Paper citation
- 3) Citation of a journal or conference
- 4) Presence of identifier (ACL, Pibmed, DOI, arXiv)
- 5) Author overlapping
- 6) Cosine similar titles of the original paper and its reference paper:
 - TF-IDF
 - W2V
 - LDA
 - Rouge scores
- 7) Topic similarity of Kullback-Leibler divergence between reference paper and other papers from collection

Ranking of papers in the scenario

Ranking of papers in the scenario
Is given: paper overview, references of paper overview
To find: order of papers by their authority
Criterion: Rank correlation

Models:

- Baseline model with $\tau = 0.1$
- rankingSVM with the pairwise transform with $\tau = 0.6$

Kendall correlation coefficient

Let (x_i, y_i) - a set of observations of the joint random variables X and Y respectively, such that all the values of (x_i) and (y_i) are unique.

Pairs (x_i, y_i) and (x_j, y_j) where $i < j$:

- Concordant if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$
- Discordant if both $x_i > x_j$ and $y_i < y_j$; or if both $x_i < x_j$ and $y_i > y_j$
- If $x_i = x_j$ and $y_i = y_j$, the pair is neither concordant nor discordant

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}$$

Citation based summarization

- 1) Preprocessing annotated sets of CPs and RPs
[Jaidka, Overview of the CL-SciSumm 2016 Shared Task, 2016](#)
- 2) Computing a set of features:

Features:

- > TF-IDF cosine similarity (tfidf)
- > latent semantic indexing (lsi) cosine similarity
- > number of common bigrams (bigrams)
- > positional features
 - position of the sentence in the RP (sid_pos)
 - position of the sentence in the section of the RP (ssid_pos)
 - position of the section in the RP (sect_pos)

Features:

- > W2V cosine similarity (w2v)
- > Word Mover's Distance between embedded word vectors (wmd)
- > Sequence Matcher (seq)
- > Rouge scores
- > Latent Dirichlet allocation cosine similarity (lda)
- > Hierarchical Dirichlet Process cosine similarity (hdp)

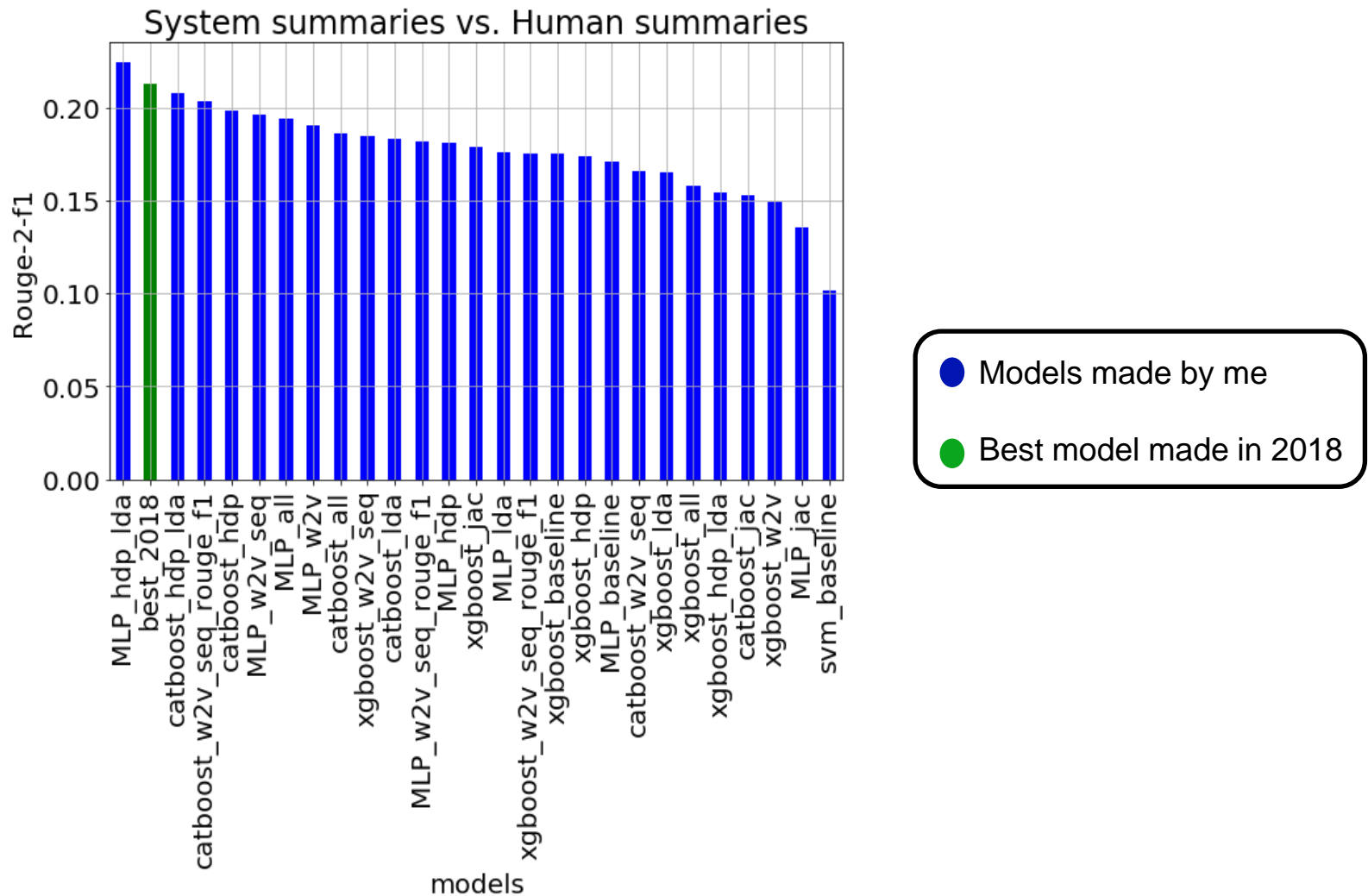
Citation based summarization

Reference span is a sentence in Reference paper which is mostly cited

- 3) **Training** any **classifier** with the goal of predicting if a sentence of the reference paper is a reference span or not
 - > Random Forest > SVM > XGBoost > CatBoost > MLP
- 4) **Summarization: Ranking** by probability sentences of references paper and selecting with the highest score for summary
 - 1 summary (total system)
 - top-k ranking summaries (system+human)
- 5) **Evaluation** by Rouge metrics

$$Rouge_n = \frac{\text{number of overlapping } n\text{-grams (human}_{summary}, \text{system}_{summary})}{\text{number of } n\text{-grams in human summary}}$$

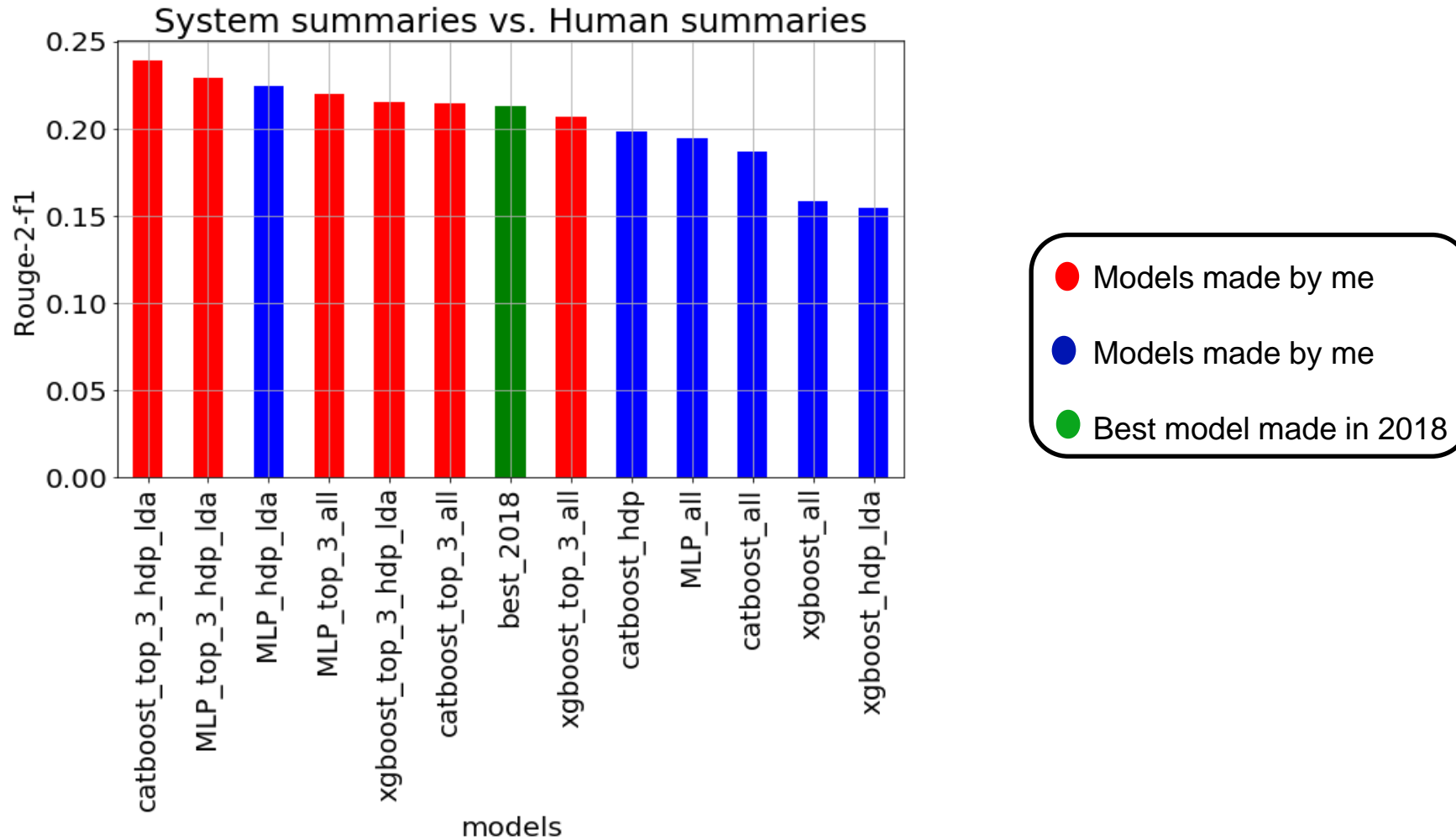
Results for **total system** summaries



Discussion of results for **total system** summaries

- best features (made by me):
 - w2v
 - wmd
 - lda
 - seq_match
- best models:
 - multilayer perceptron
 - catboost
- our summarization model works better than the **best_2018** model

Results for **total system** & **system+human** summaries



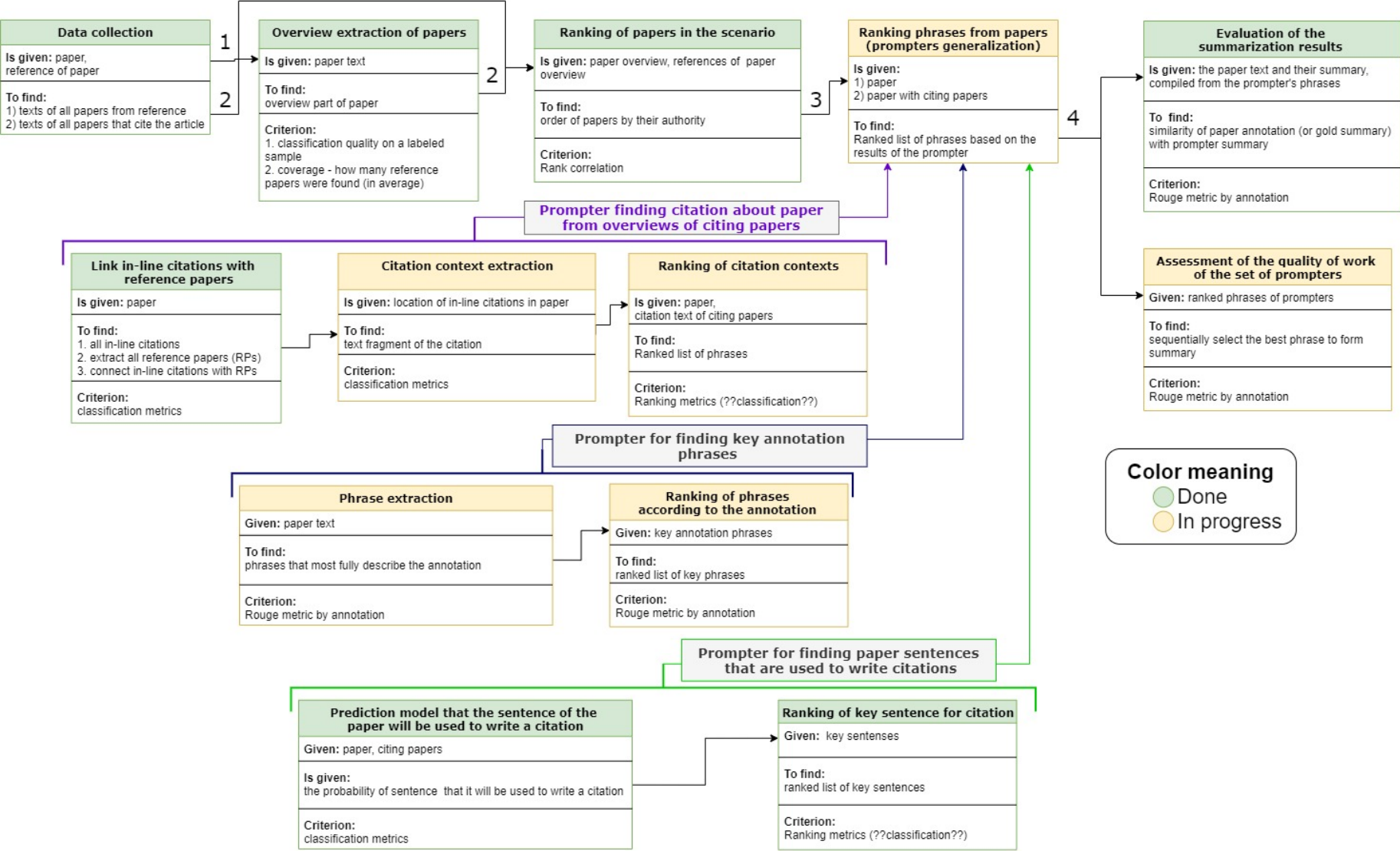
Discussion of results for **system+human** summaries

- Rouge metric for **system+human** summaries are 15-19% better than for **total system** summaries
- the best **total system** classifier == the best **system+human** classifier

Conclusions

1. New approach for generating background section was developed
2. The whole work was made from the beginning to the end (more summarization methods will be realized in the future)
3. The citation based summarization from reference achieves excellent results.

Current Status



Outlook

1. To improve achieved results
2. To add new prompters
3. To implement our solution to <https://arxiv-search.mipt.ru/>

Machine-aided multi-document summarization of scientific papers

Student: *Andrey Vlasov*

Research Advisor: *Maxim Panov*

Co-Advisor : *Konstantin Vorontsov*

May, 2020