Data Science

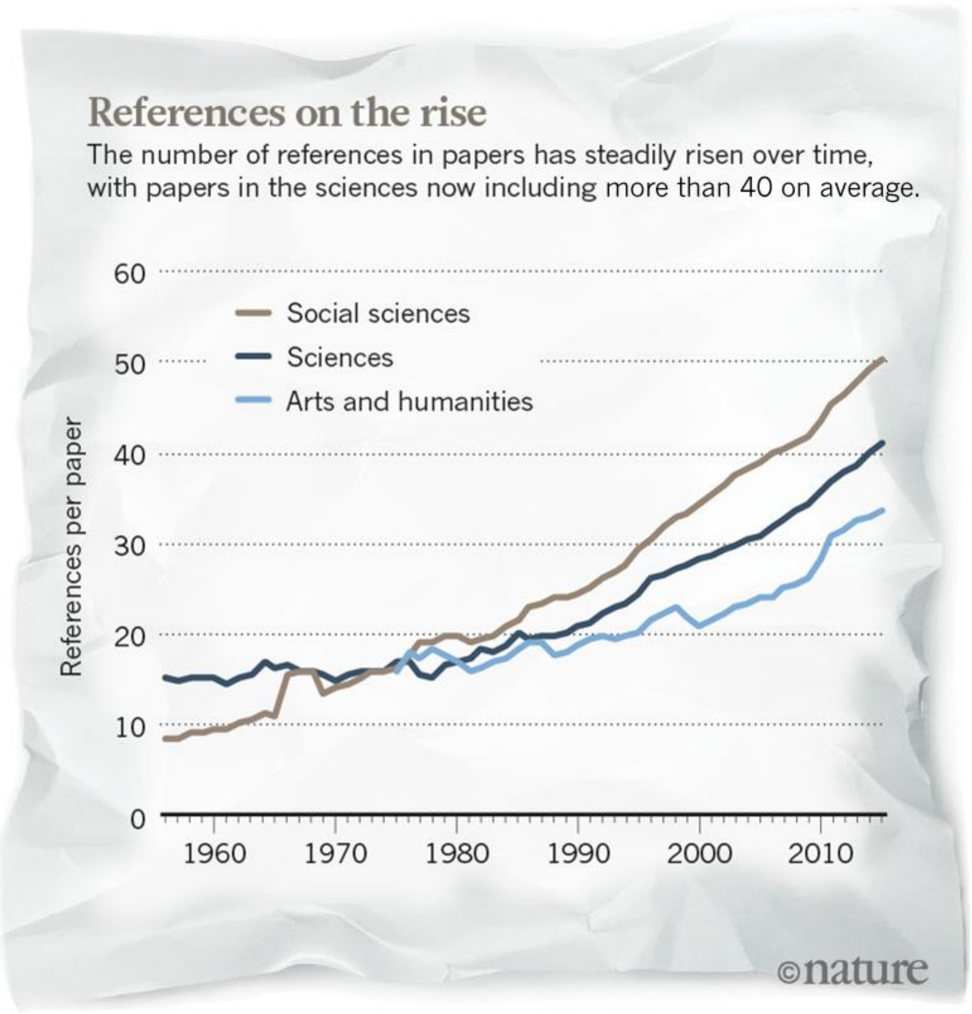
Machine-aided multi-document summarization of scientific papers

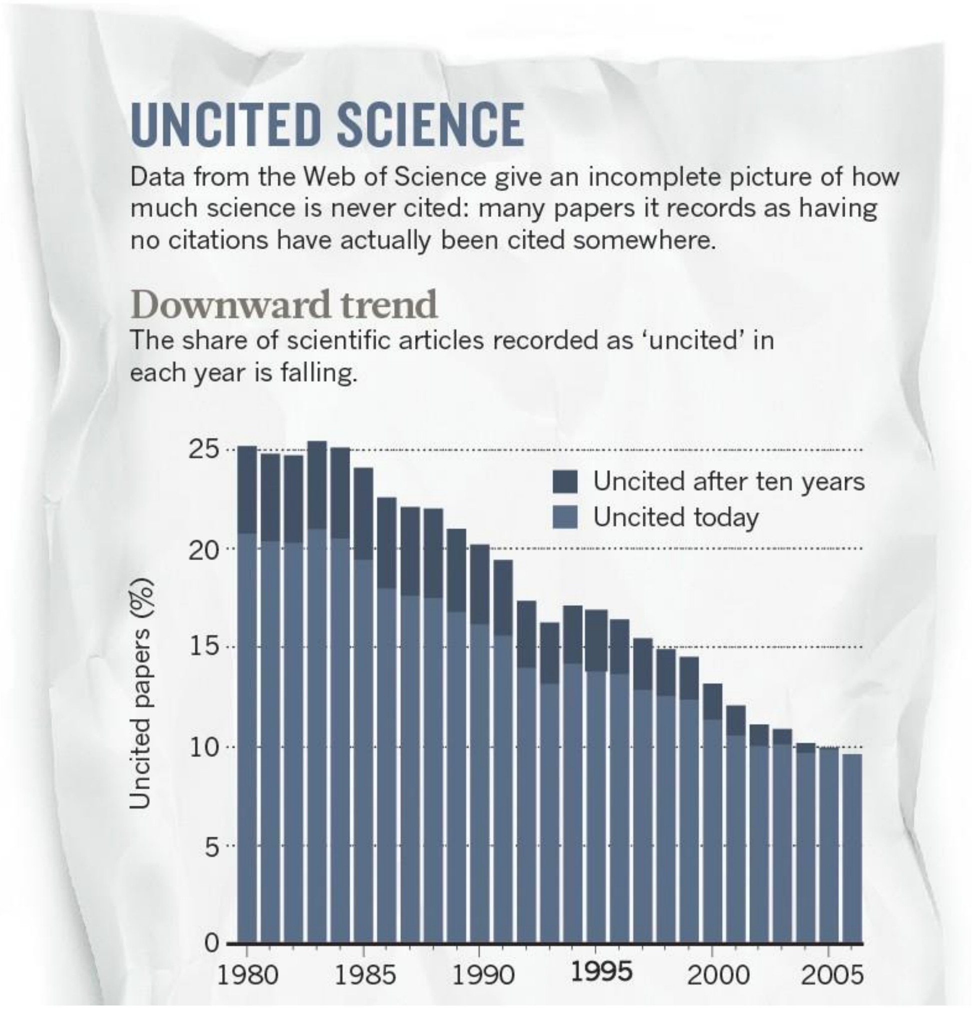
Student: *Andrey Vlasov*
Research Advisor: *Maxim Panov*
Co-Advisor : Konstantin Vorontsov

May, 2020

Skoltech

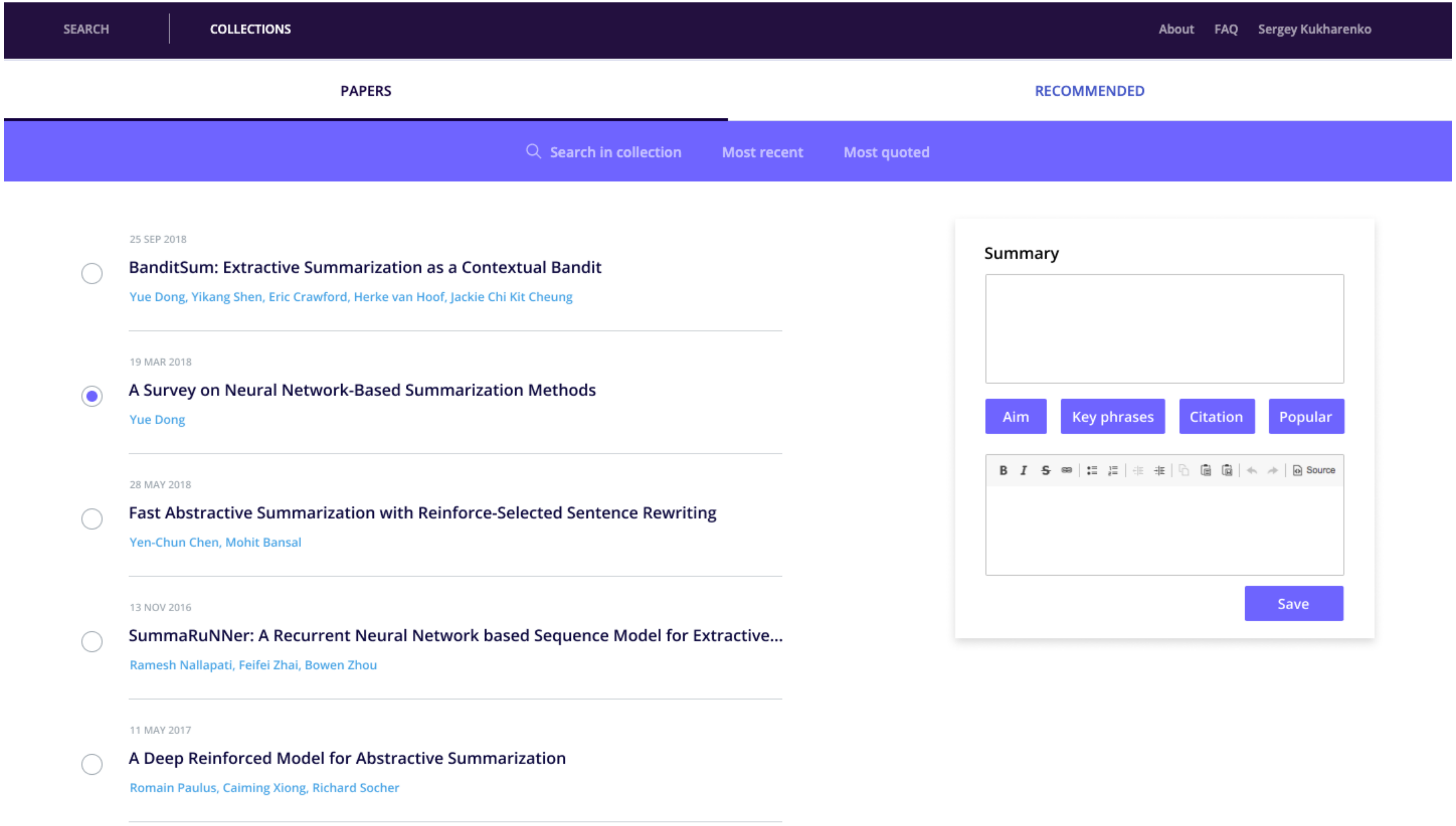
Stating the problem



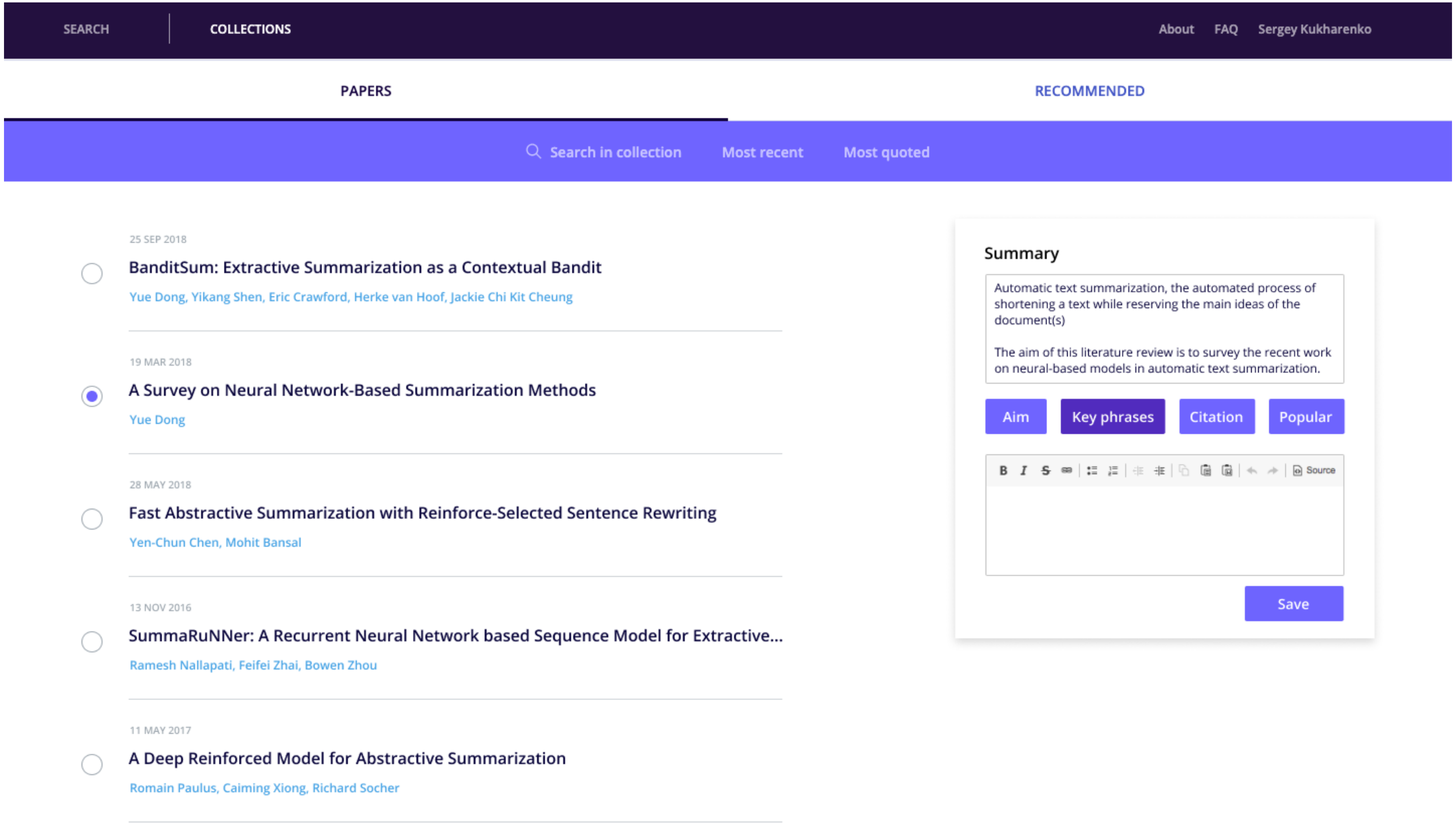


Skoltech

Introduction



Introduction



Aim

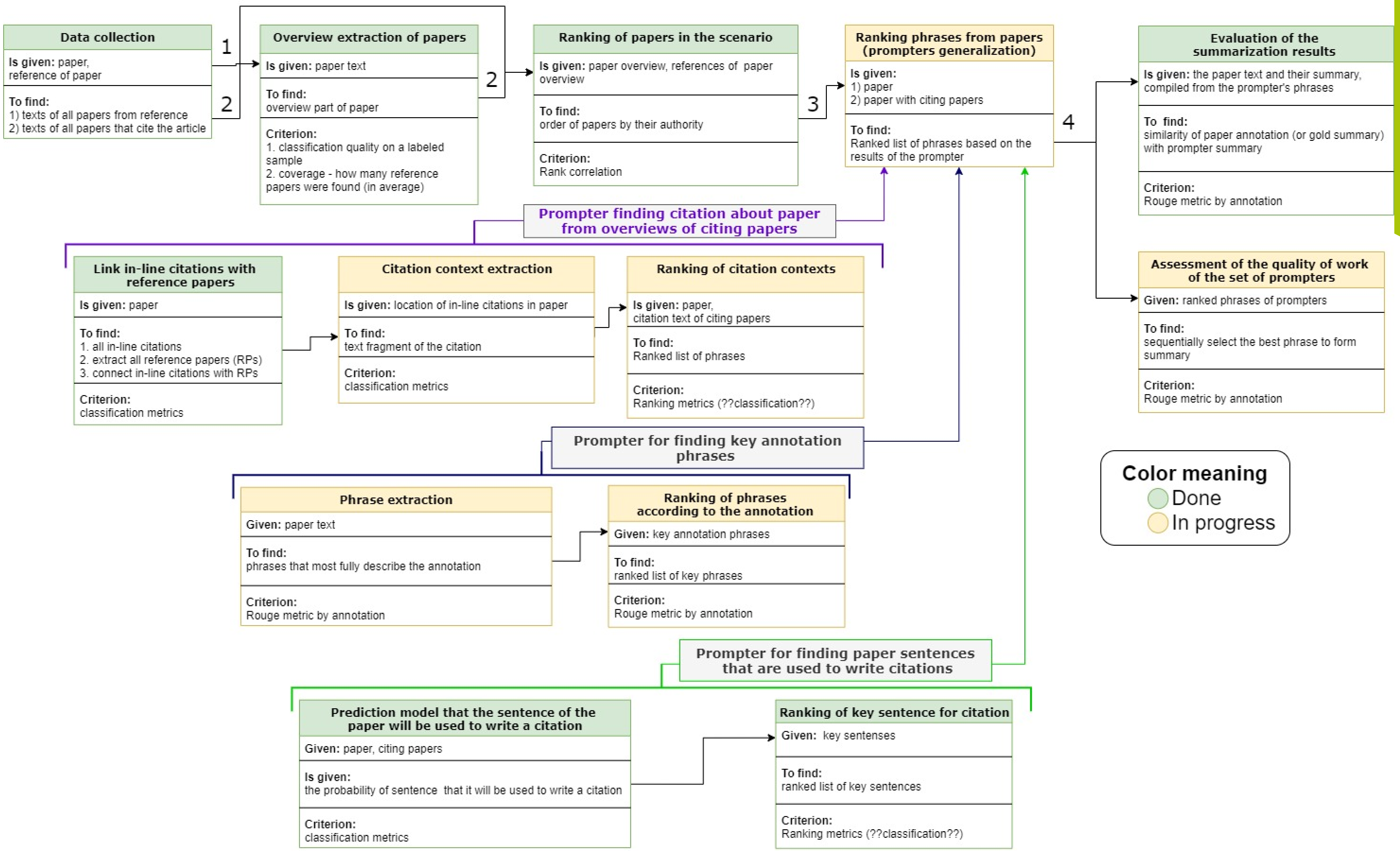
To create and implement a methodology for solving task of automated multi-document summarization of scientific papers

Objectives

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

Skoltech

Pipeline



Skoltech

Data collection

[S2ORC: The Semantic Scholar Open Research Corpus]

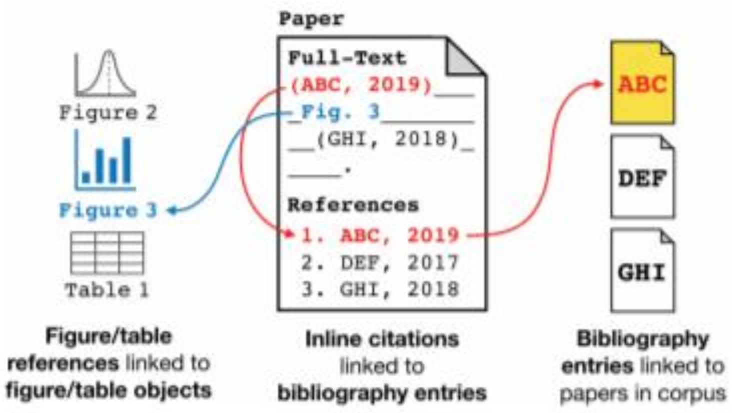


Figure 1. Inline citations and references are annotated in full text, bibliography entries, and figure and table captions are preserved; citations are linked to bibliography entries, which are linked to other papers in S2ORC.

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%)

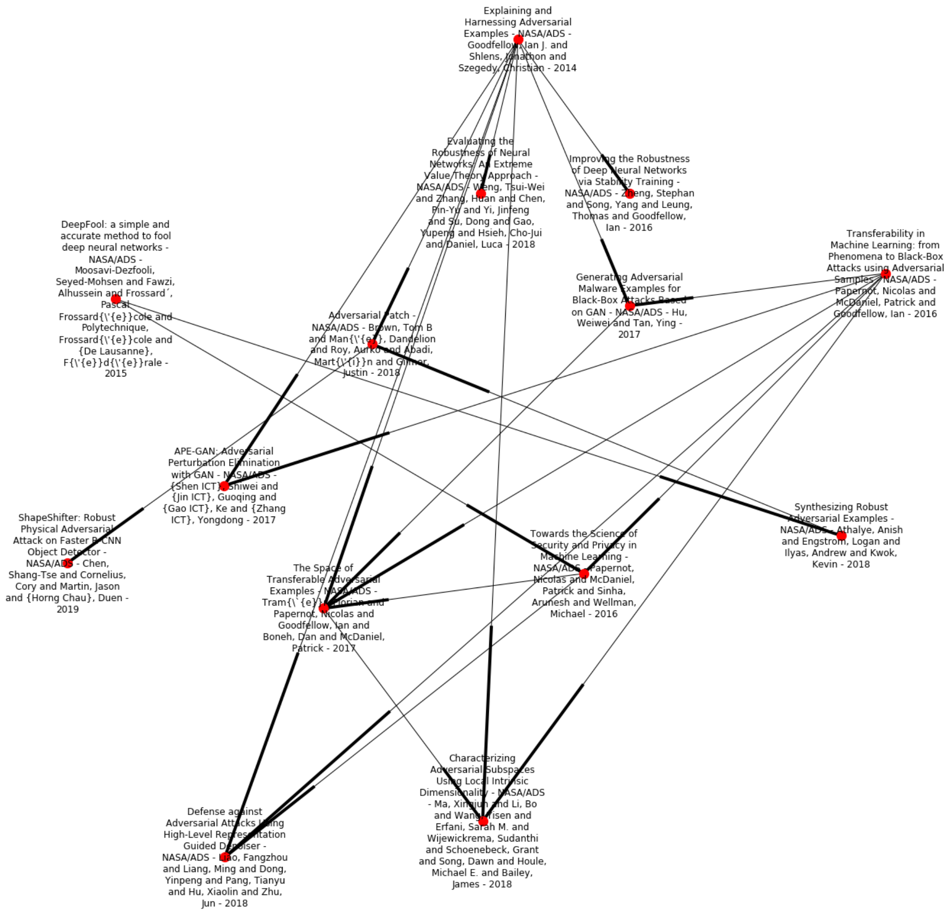
Table 1. Statistics of papers in this dataset

Skoltech

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

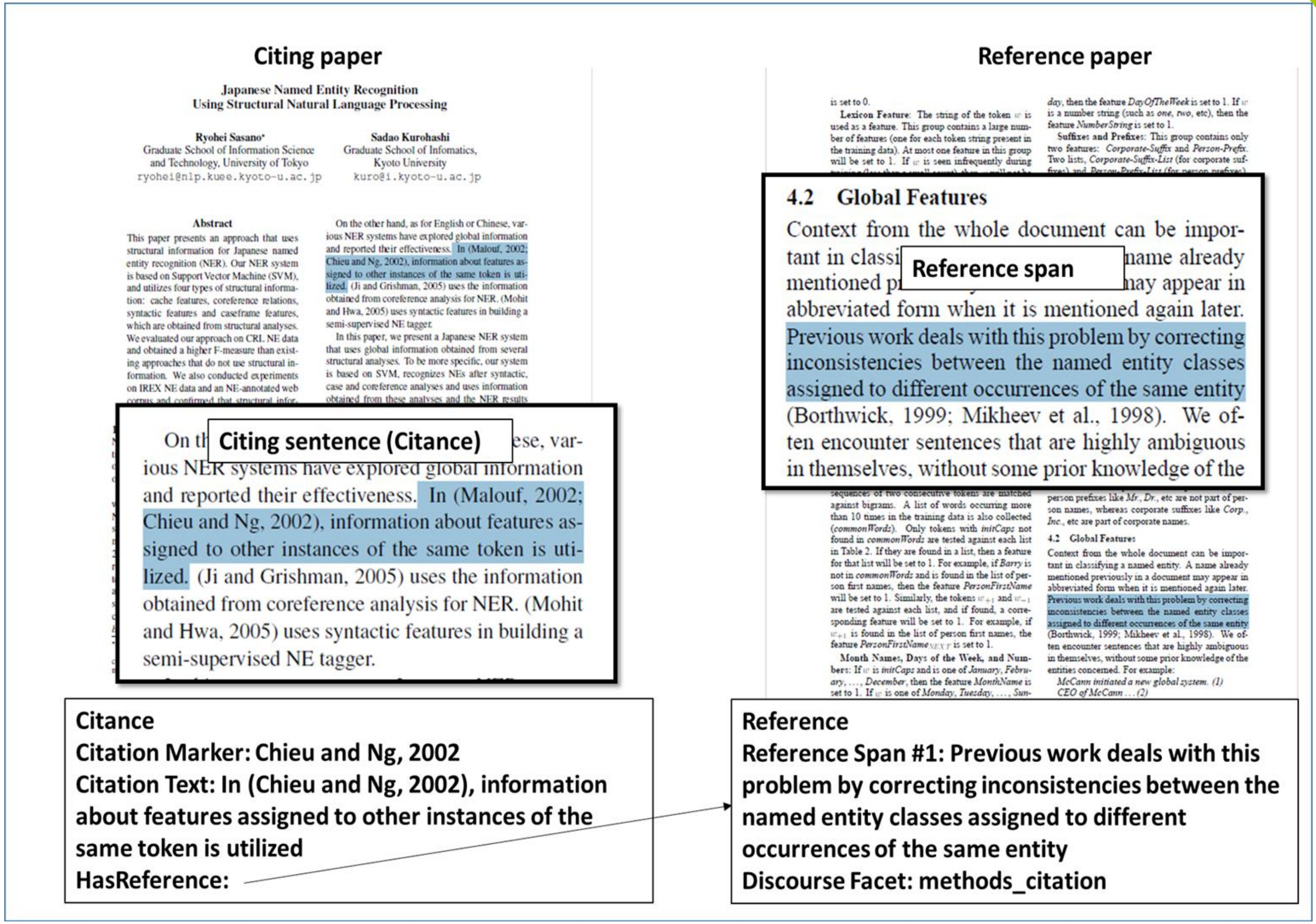


Skoltech

Dataset

[CL-SciSumm]

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



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 t0 = 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

Skoltech

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 <u>features</u>:

- 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

<u>Results</u>:

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

Skoltech

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 τ = *0.1*
- rankingSVM with the pairwise transform with τ = *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 $\left(x_j, y_j\right)$ 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 $x_i = x_i$ and $y_i = y_i$, 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)

Skoltech

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

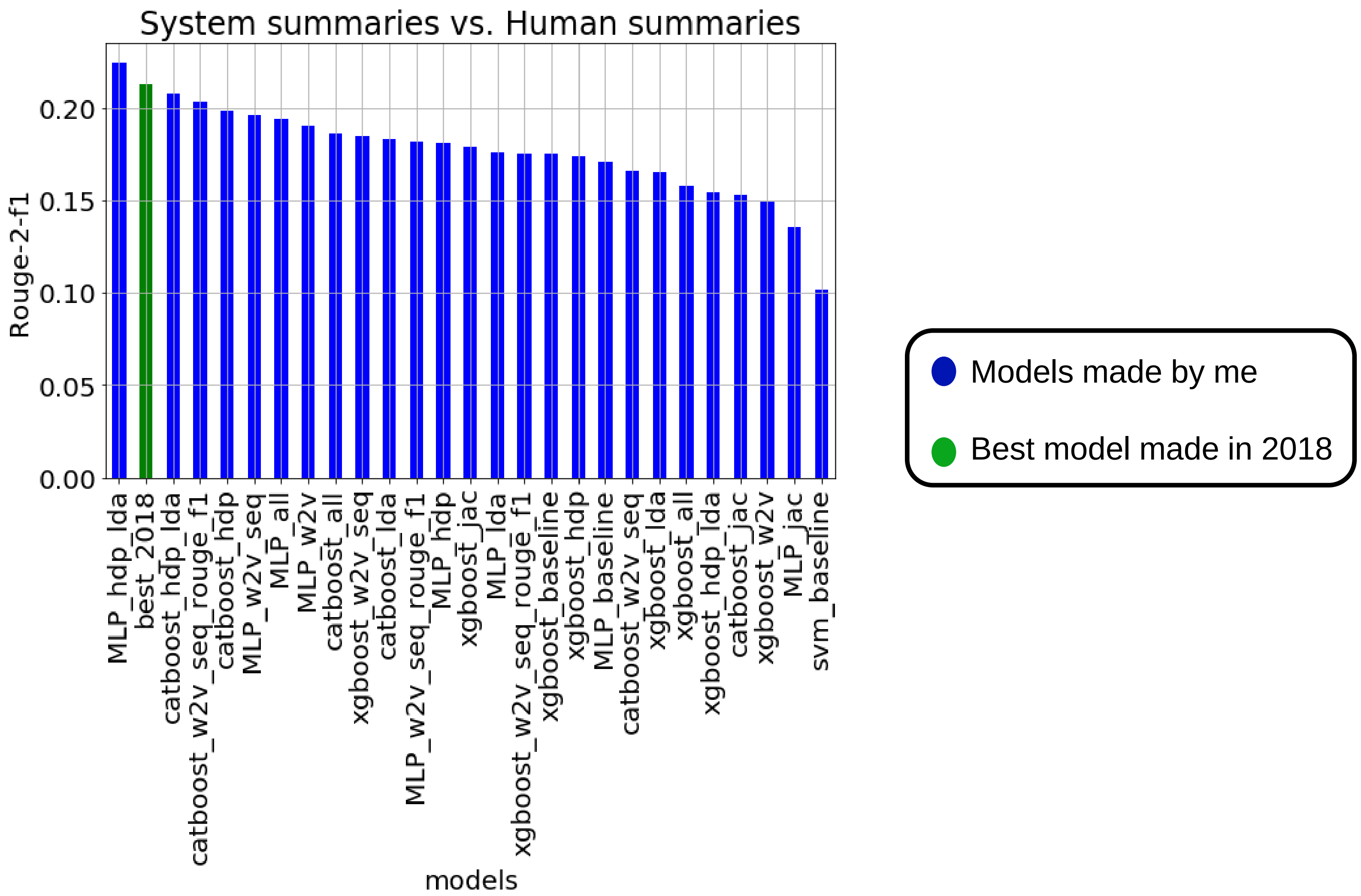
4) **Summarization: Ranking** by probability sentences of references paper and selecting with the highest score for summary
   - ➢ 1 summary (total system)
   - ➢ top-k ranking summaries (system+human)

5) **Evaluation** by Rouge metrics

$$Rouge_n = \frac{number\ of\ overlapping\ n\!-\!grams\ (human_{summary}, system_{summary})}{number\ of\ n\!-\!grams\ in\ human\ summary}$$

Skoltech

Results for **total system** summaries

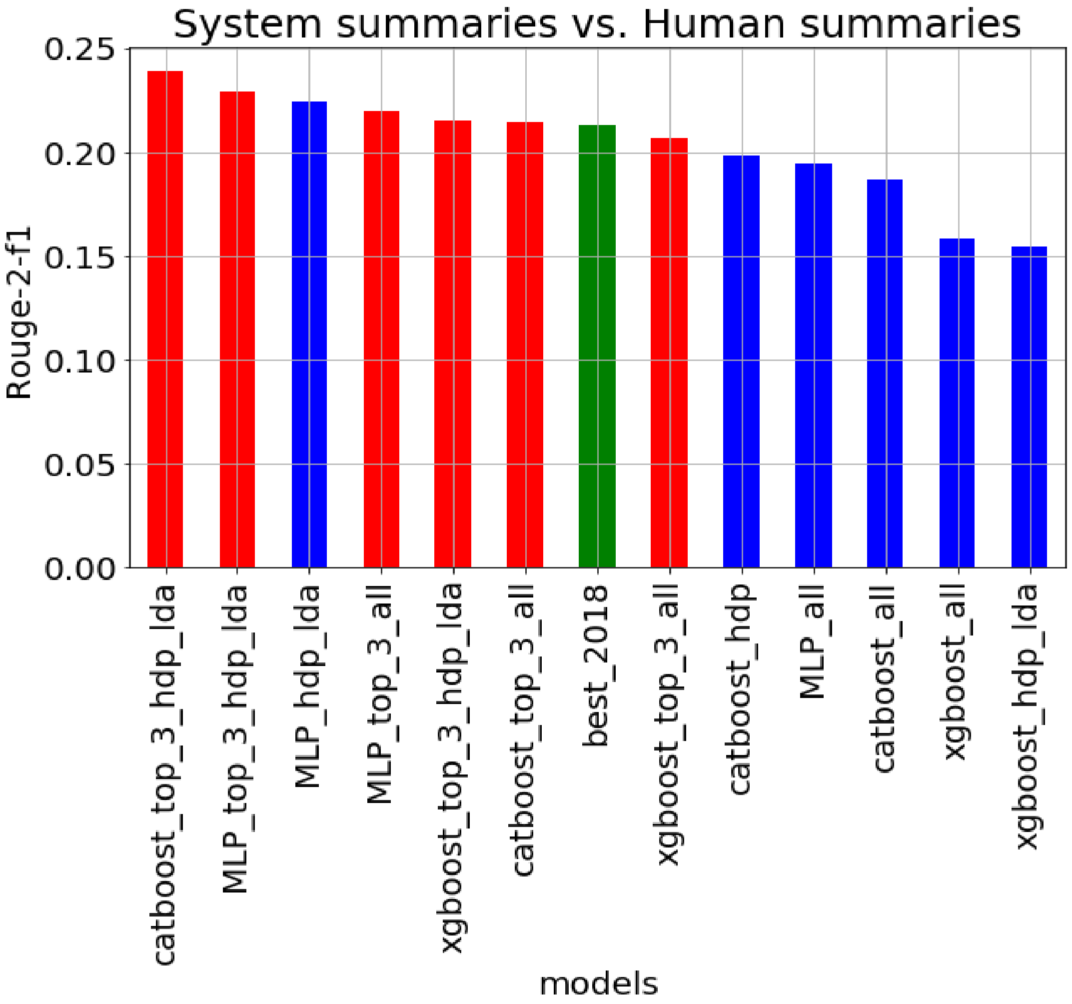


Discussion of results for total system summaries

- 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

Skoltech

Results for total system & system+human summaries



- Models made by me
- Models made by me
- Best model made in 2018

Skoltech

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

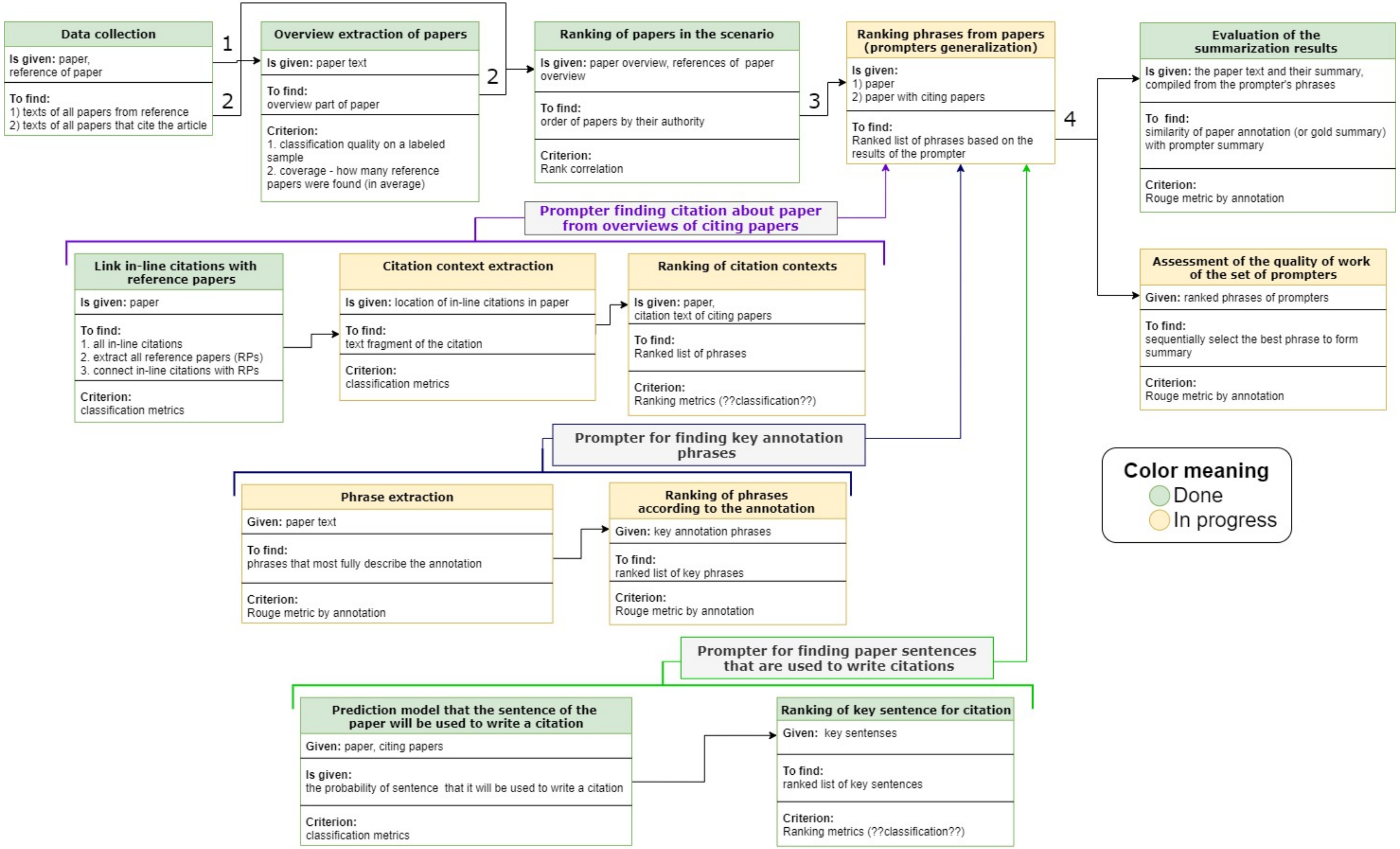
Skoltech

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.

Skoltech

Current Status



Skoltech

Outlook

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

Data Science

Machine-aided multi-document summarization of scientific papers

Student: *Andrey Vlasov*
Research Advisor: *Maxim Panov*
Co-Advisor : Konstantin Vorontsov

May, 2020

Skoltech