*Data Science*

**Machine-aided multi-document summarization of scientific papers**

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

May, 2020

**Stating the problem**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 2

**Introduction**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 3

**Introduction**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 4

**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)

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 5

**Pipeline**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 6

**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.

Table 1. Statistics of papers in this dataset

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 7

S2ORC: The Semantic Scholar Open Research Corpus

**Data collection**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 8

**Dataset**

CL-SciSumm

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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 9

**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**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 10

**The extraction of an overview part**

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)

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 11

**Ranking of papers from the collection**

Features: 1) Year of publication 2) Paper citation 3) Сitation 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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 12

**Ranking of papers in the scenario**

**Models**:

▪ Baseline model with *τ = 0.1*

▪ rankingSVM with the pairwise transform with *τ = 0.6*

**Kendall correlation coefficient** Let xi,yi - a set of observations of the joint random variables X and Y respectively, such that all the values of xi and yi are unique. Pairs xi,yi and xj,yj where i < j: - Concordant if both xi > xj and yi > yj ; or if both xi < xj and yi < yj - Discordant if both xi > xj and yi < yj ; or if both xi < xj and yi > yj - If xi = xj and yi = yj , the pair is neither concordant nor discordant

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 13

**Citation based summarization**

1) Preprocessing annotated sets of CPs and RPs 2) Jaidka, Computing Overview a set of of features:

the CL-SciSumm 2016 Shared Task, 2016 **Features:**

**Features:** > TF-IDF cosine similarity (tfidf)

> W2V cosine similarity (w2v) > latent semantic indexing (lsi) cosine similarity

> Word Mover’s Distance between embedded word > number of common bigrams (bigrams)

vectors (wmd) > positional features - position of the sentence in the RP (sid\_pos) - position of the sentence in the section of the

> Sequence Matcher (seq) > Rouge scores > Latent Dirichlet allocation cosine similarity (lda) RP (ssid\_pos) - position of the section in the RP (sect\_pos)

> Hierarchical Dirichlet Process cosine similarity (hdp)

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 14

**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

Rougen = number of overlapping n−grams (humansummary,systemsummary)

number of n−grams in human summary

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 15

**Results for total system summaries**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 16

Models made by me

Best model made in 2018

**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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 17

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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 18

Models made by me

Models made by me

Best model made in 2018

**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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 19

**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.

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 20

**Current Status**

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 21

**Outlook**

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

Andrey Vlasov. Machine-aided multi-document summarization of scientific papers 22

*Data Science*

**Machine-aided multi-document summarization of scientific papers**

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

May, 2020