Marco Polo

Testing retrieval performance on summarized documents

Group 25, `github.com/muehlt/marco-polo`

Fabian Staber

Analysis, Plotting

Matthias Paltauf

Evaluation Pipeline

Martin Brantner

Document summarization

Thomas Mühlbacher

Data Processing, Word2Vec

Research Question

What effect does document summarization have on Word2Vec based retrieval performance?

Motivation

- **Save space** if smaller summarizations are sufficient for searches
- **Evaluate** practical summarization performance
- **Evaluate** amount of data needed for Word2Vec training

Dataset

Microsoft

MS MARCO

https://microsoft.github.io/msmarco/

corpus.jsonl

Web Document Passages

8.8 Mio

```jsonl
{"_id": "0", "title": "", "text": "The presence of communication amid scientific minds was equally i
{"_id": "1", "title": "", "text": "The Manhattan Project and its atomic bomb helped bring an end to
{"_id": "2", "title": "", "text": "Essay on The Manhattan Project - The Manhattan Project The Manhat
{"_id": "3", "title": "", "text": "The Manhattan Project was the name for a project conducted during
{"_id": "4", "title": "", "text": "versions of each volume as well as complementary websites. The fi
{"_id": "5", "title": "", "text": "The Manhattan Project. This once classified photograph features t
```

query.jsonl

Real User Queries

509.963

```jsonl
{"_id": "1185869", "text": ")what was the immediate impact of the success of the manhattan project?"
{"_id": "1185868", "text": "_____ justice is designed to repair the harm to victim, the communit
{"_id": "597651", "text": "what color is amber urine", "metadata": {}}
{"_id": "403613", "text": "is autoimmune hepatitis a bile acid synthesis disorder", "metadata": {}}
{"_id": "1183785", "text": "elegxo meaning", "metadata": {}}
{"_id": "312651", "text": "how much does an average person make for tutoring", "metadata": {}}
```

test.tsv

Graded Relevance Judgements

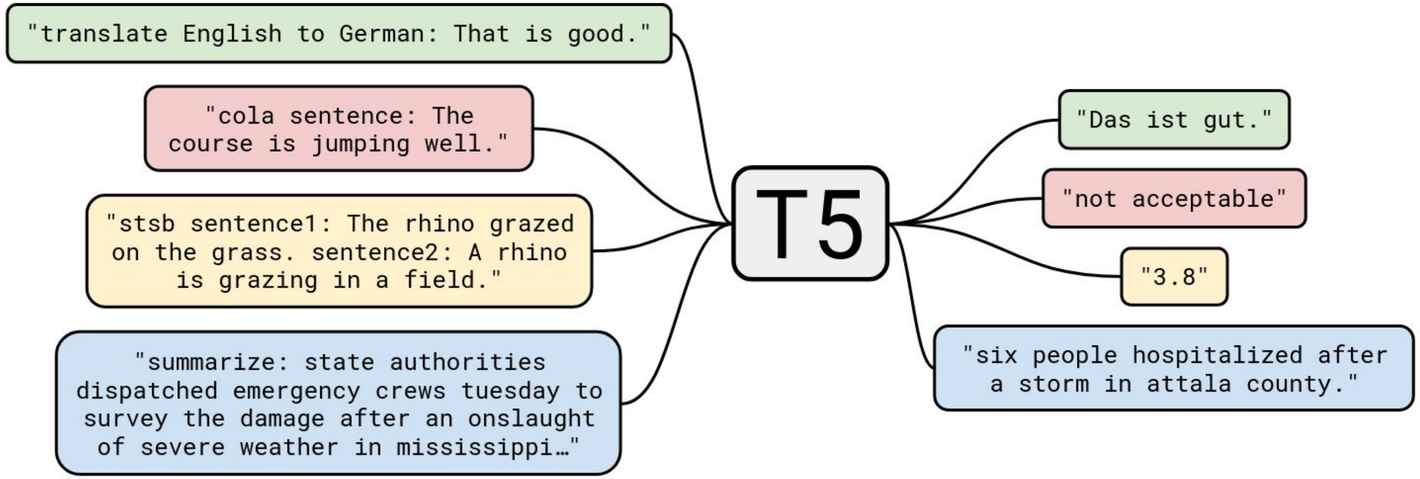
9.262

query-id	corpus-id	score
19335	1017759	0
19335	1720389	1
19335	1729	2
19335	1730	0
19335	1796642	0

Summarization

Text-**T**o-**T**ext **T**ransfer

Transformer Model



https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

Model checkpoint

T5-Small

Parameters

60 Mio

Average summarization runtime

30 Min

Capabilities

- Classification
- Summarization
- Question answering
- Translation

Dataset processing

Preparation Steps

Fetching

Download data from repo, if necessary

Extract files for local usage

Reducing

Reduce corpus & queries to match available relevance scores

Train Word2Vec model on reduced corpus, queries & summaries (+ optionally non-used document words)

Cleaning

Remove irrelevant characters, like unicode hex sequences or HTML encodings

Data pairs without useful (tokenizable) remaining data **filtered during runtime of evaluation**

Dataset processing

During Execution

Loading

Load data files for evaluation

corpus.reduced.jsonl

queries.reduced.jsonl

summaries.jsonl

Load relevance scores

qrels/test.tsv

Processing

Remove punctuation

Remove stopwords

Tokenize words

Stem words

Word2Vec

Embed words using Word2Vec model

Get document vectors by averaging

Retrieval

Cosine Similarity

Queries

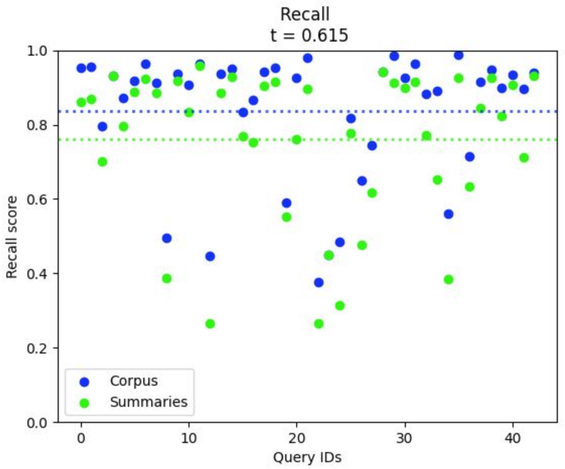
Document ids

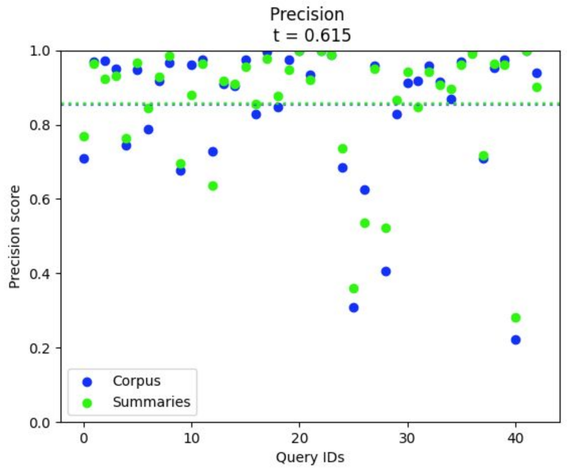
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
1017759	0.79243	0.45887	0.66988	0.52182	0.48655	0.62023	0.66082	0.51797	0.57099	0.61566	0.65155	0.31664	0.62467	0.83206
1082489	0.88629	0.56196	0.38554	0.56421	0.74434	0.65859	0.75079	0.82181	0.73189	0.53751	0.56762	0.50815	0.86527	0.55101
109063	0.93014	0.53517	0.34239	0.64758	0.78361	0.66851	0.84507	0.71897	0.69471	0.60638	0.72025	0.39707	0.86211	0.60980
1160863	0.90258	0.51457	0.46388	0.48771	0.57243	0.48731	0.71309	0.58620	0.70487	0.60260	0.61378	0.24921	0.73901	0.64076
1160871	0.90434	0.51942	0.46289	0.48484	0.57849	0.49216	0.71219	0.59592	0.70886	0.59945	0.61135	0.25530	0.74403	0.63635
1189088	0.91160	0.47324	0.48047	0.54604	0.63130	0.57665	0.73992	0.57469	0.63542	0.55804	0.71208	0.37029	0.75119	0.71082
1203500	0.80112	0.61537	0.46401	0.63537	0.77014	0.56875	0.78605	0.73955	0.69657	0.65773	0.73991	0.30088	0.79478	0.61348
1231806	0.87427	0.44263	0.39742	0.54414	0.76561	0.73369	0.74712	0.69124	0.53845	0.52828	0.75089	0.37209	0.73354	0.60956
1231807	0.80958	0.35516	0.49584	0.44778	0.68182	0.85451	0.60938	0.62292	0.57001	0.57800	0.66357	0.48914	0.69535	0.61250
1274615	0.81219	0.63520	0.59342	0.58867	0.57127	0.62163	0.72220	0.66960	0.68060	0.56061	0.56169	0.45210	0.74893	0.76828
1274620	0.84113	0.52998	0.59482	0.54116	0.51177	0.48598	0.67185	0.56780	0.65096	0.56191	0.58950	0.27782	0.68310	0.80722
1324075	0.50893	0.71315	0.60073	0.45577	0.42346	0.46946	0.54728	0.61066	0.66829	0.49067	0.36906	0.68094	0.64117	0.51446
1509459	0.87106	0.42724	0.49699	0.38005	0.45697	0.47207	0.59182	0.56396	0.60931	0.41872	0.46929	0.34775	0.66491	0.67477
1555317	0.87922	0.57748	0.39684	0.60969	0.73940	0.61553	0.81783	0.70309	0.70131	0.65607	0.70451	0.24119	0.81340	0.61985
1568085	0.97389	0.43771	0.24081	0.60520	0.80224	0.61104	0.84029	0.72735	0.63647	0.50267	0.66282	0.31742	0.85288	0.54250
161603	0.83691	0.59117	0.49355	0.55199	0.73635	0.64910	0.72366	0.76895	0.64376	0.53030	0.66701	0.40366	0.77298	0.64969
1705525	0.91740	0.39625	0.31980	0.53346	0.79554	0.79532	0.74744	0.72734	0.58119	0.50535	0.67863	0.45045	0.79641	0.54026
1720387	0.80792	0.67408	0.59483	0.57425	0.56853	0.45282	0.70090	0.59785	0.80374	0.69325	0.63799	0.31093	0.75969	0.67633
1720388	0.90424	0.55323	0.49445	0.57157	0.66560	0.56808	0.71121	0.68055	0.74864	0.61589	0.64805	0.35835	0.80549	0.67214
1720389	0.82349	0.63221	0.56770	0.59411	0.59321	0.48747	0.69980	0.57670	0.80643	0.72170	0.67393	0.28181	0.77062	0.67872
1720393	0.83823	0.50949	0.60499	0.54378	0.58202	0.62064	0.70166	0.56323	0.58776	0.55944	0.70221	0.41431	0.68713	0.78459
1720395	0.82409	0.64145	0.53961	0.57994	0.59023	0.45652	0.69647	0.57863	0.82516	0.71725	0.65083	0.25832	0.77712	0.64368

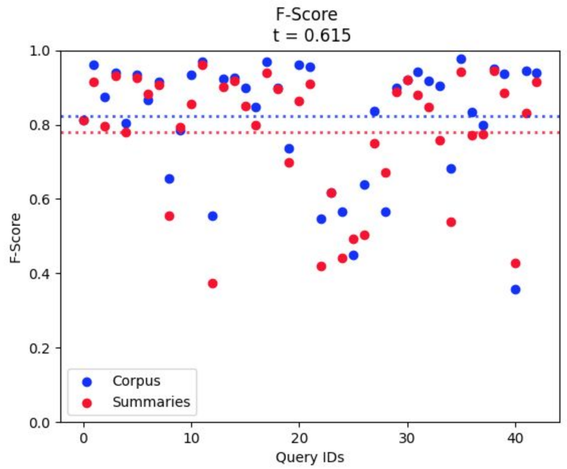
Result

Evaluation Metrics

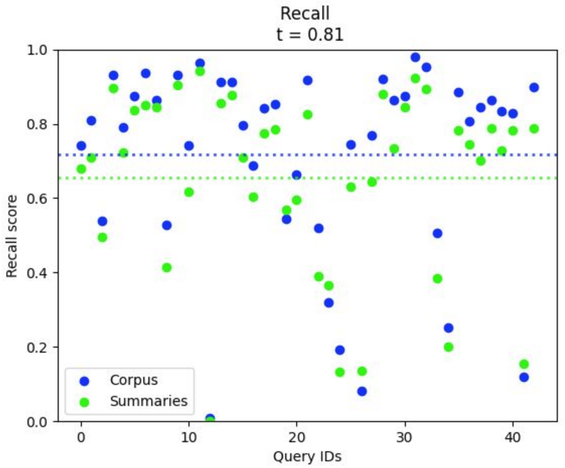
Skip-gram

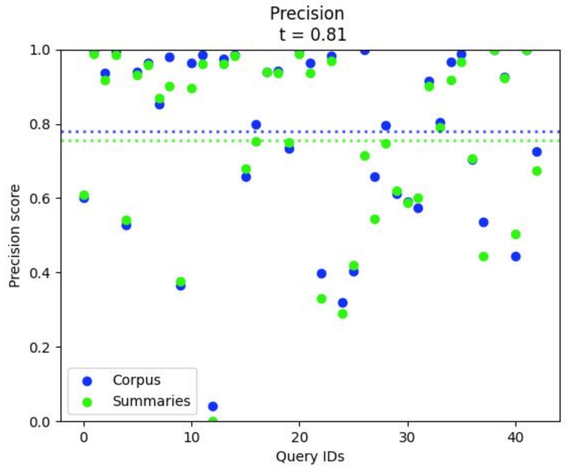


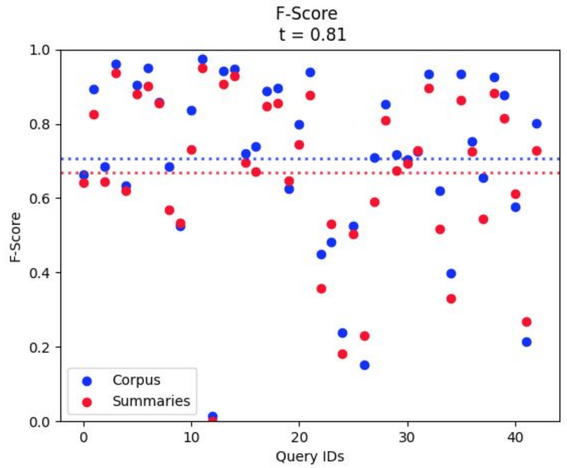




CBOW



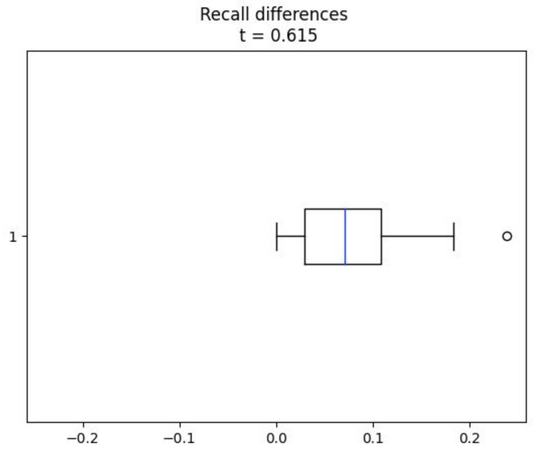


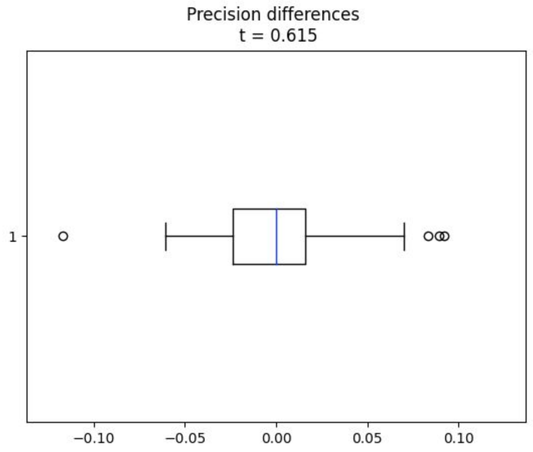


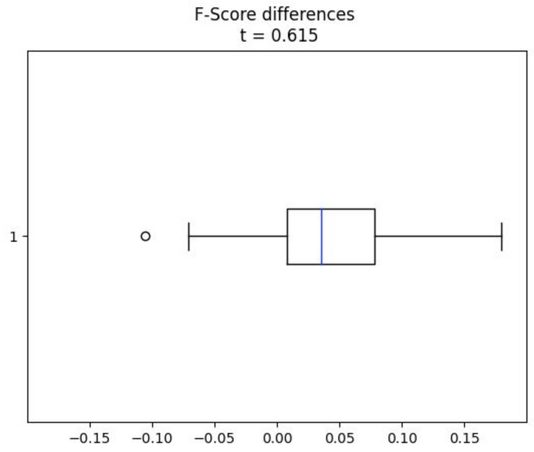
Result

Evaluation Metrics

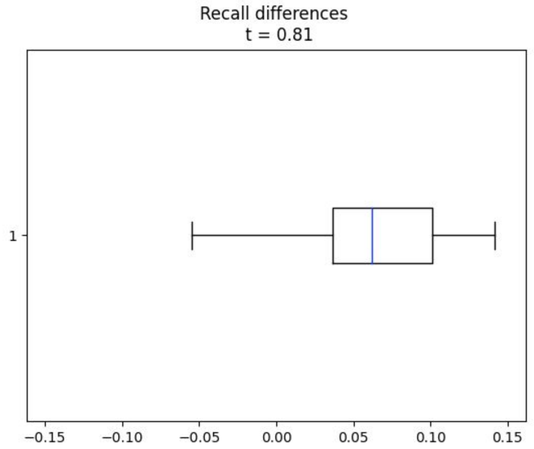
Skip-gram

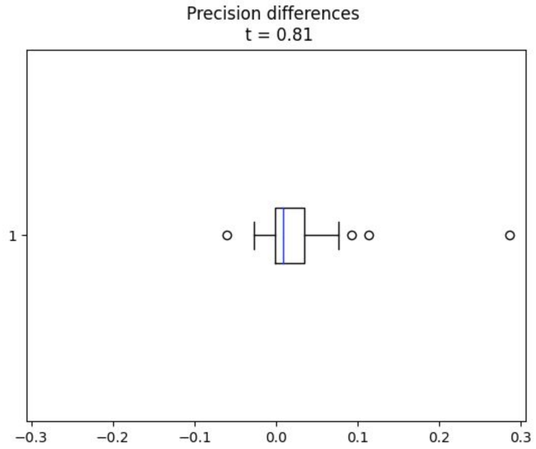


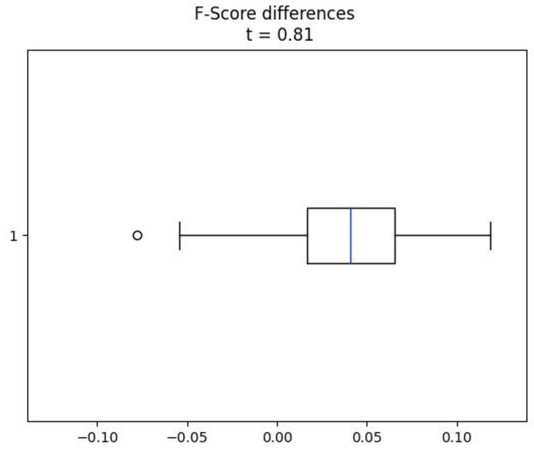




CBOW







Conclusion

Learnings:

Worse but sufficient retrieval performance for summarized docs

Use Skip-gram for small datasets

Unextended, reduced data sufficient for Word2Vec training

Limitation:

T5 performance

Bias:

Data only contains short, english texts

Thank you for listening.

We are happy to answer any questions about the project.

GitHub Repository

github.com/muehlt/marco-polo

Dataset MS Marco

microsoft.github.io/msmarco/