CLEF 2024 GRENOBLE

SIMPLE TEXT

09-09-2024

Enhancing Scientific Document Simplification through Adaptive Retrieval and Generative Models

Artemis Capari, Hosein Azarbonyad, Zubair Afzal, Georgios Tsatsaronis

ELSEVIER

TABLE OF CONTENTS

01

Task 1: ***What's in or out?***

→ Retrieve all passages pertinent to a given query or topic

- 2023: Improve ranking model for scientific passage retrieval task
- 2024: Improve input to ranking models by generating new search queries

Task 3: ***Rewrite this!***

→ Simplify passages from scientific abstracts given a query

- Test power of simple prompt engineering

Introduction

02

TASK 1: *What's in or out?*

Retrieve all passages pertinent to a given query or topic

02 TASK 1 2023

Fine-tuning dense-retrieval models

- Validation:
  - 100 queries across 20 disciplines.
  - Pooling technique for document retrieval
  - 50 manually evaluated snippets per query
- Training:
  - Large set of unlabeled documents
  - Generative Pseudo Labeling (GPL)
    - Unsupervised domain adaptation
    - Generates pseudo labels for unlabeled data
    - Pseudo labels with ms-marco-MiniLM-L-6-v2

02 TASK 1 2023

Table: Details on fine-tuning of various models

Model Name	Bi-Encoder	Queries	Documents	Batch Size	Training Steps	Epochs
MS-DB-v4-GPL-CS	msmarco-distilbert-base-v4	218 (10 golden)	23670	16	15000	1
MS-DB-tas-b-GPL-CS	msmarco-distilbert-base-tas-b	218 (10 golden)	23670	16	15000	1
MS-DB-v4-GPL-all	msmarco-distilbert-base-v4	4637 (80 golden)	893110	32	280000	1
MS-DB-tas-b-GPL-all	msmarco-distilbert-base-tas-b	4637 (80 golden)	893110	32	280000	1

02 — TASK 1 — 2023

Table: Official Results of Simple Text Task 1 - CLEF 2023

Run	MRR	P@10	P@20	P@30	nDCG@10	nDCG@20	nDCG@30	BPREF	MAP
ElsevierSimpleText_run8	0.8082	0.5618	0.3515	0.2696	0.5881	0.4422	0.3803	0.2371	0.1633
ElsevierSimpleText_run7	0.7136	0.5618	0.4103	0.3441	0.5704	0.4627	0.4158	0.2626	0.1915
maine_CrossEncoder1	0.7309	0.5265	0.4500	0.4216	0.5455	0.4841	0.4687	0.3337	0.2754
maine_CrossEncoderFinetuned1	0.7338	0.4971	0.4000	0.3529	0.4859	0.4295	0.4062	0.3443	0.2385
ElsevierSimpleText_run5	0.6600	0.4765	0.3838	0.3314	0.4826	0.4186	0.3834	0.2542	0.1828
ElsevierSimpleText_run2	0.7010	0.4676	0.4059	0.3480	0.4791	0.4282	0.3912	0.2528	0.1942
ElsevierSimpleText_run6	0.6402	0.4676	0.3853	0.3284	0.4723	0.4185	0.3828	0.2557	0.1809
ElsevierSimpleText_run4	0.6774	0.4529	0.3794	0.3422	0.4721	0.4116	0.3876	0.2485	0.1898
ElsevierSimpleText_run9	0.5933	0.4735	0.3176	0.2500	0.4655	0.3595	0.3102	0.1758	0.1238
ElsevierSimpleText_run1	0.6821	0.4588	0.3824	0.3353	0.4626	0.4071	0.3786	0.2573	0.1823
maine_CrossEncoderFinetuned2	0.7082	0.4706	0.3926	0.3637	0.4617	0.4089	0.3969	0.3259	0.2253
UAms_CE1k_Filter	0.6403	0.4765	0.3559	0.2941	0.4533	0.3743	0.3334	0.2727	0.1936
ElsevierSimpleText_run3	0.6502	0.4471	0.3779	0.3324	0.4460	0.3994	0.3709	0.2558	0.1785

...

02 TASK 1 2024

Improving input to ranking model

- Retrieving Relevant Passages
  - Need high-performing ranking model
  - How do you ask for relevant passages?
  - —> Improve search queries
- Generating Search Queries
  - GPT-3.5-turbo-0125
  - Use generated queries to:
    - Create corpus from top-k ElasticSearch documents
    - Re-rank using fine-tuned model

02 TASK 1 2024

Generating Topics using Abstracts

```
Goal:
I have a task to retrieve passages that help understand a given article.

Request:
Your task is to help me write the best possible search query to retrieve
articles that would help understand the provided article.
This query should be concise and focus on the provided topic.
Only provide ONE search query.

Article:
"{abstract}"

Search Query:
```

02

TASK 1

2024

Generate Queries using provided Topics and Abstracts

```
Goal:
I have a task to retrieve passages that help understand a given article.
We dissect the content of the article into key-topics, and retrieve
passages for those topics.

Request:
I need your help to create the best possible search query for a given
topic in the context of the provided article.
This query should be concise and focus on the provided topic.
Only provide ONE search query.

Topic:
{query_text}

Article:
"{abstract}"

Search Query:
```

02

TASK 1: Experiments

Run	Query Input	Corpus	Model
1	query	ES Top-100	MS-DB-tas-b-GPL-all
2	query, topic	ES Top-500	MS-DB-v4-GPL-CS
3	query	ES Top-1000	MS-DB-tas-b-GPL-all
4	query, topic	ES Top-100	MS-DB-tas-b-GPL-all
5	query, topic	ES Top-500	MS-DB-tas-b-GPL-all
6	query, topic	ES Top-1000	MS-DB-tas-b-GPL-all
7	query	ES Top-500	MS-DB-v4-GPL-CS
8	gen query	ES Top-100	MS-DB-tas-b-GPL-all
9	topic	ES Top-500	MS-DB-tas-b-GPL-all
10	gen topic	ES Top-100	MS-DB-tas-b-GPL-all

02 TASK 1: Experiments

Run	Query Input	Corpus	Model
1	query	ES Top-100	MS-DB-tas-b-GPL-all
2	query, topic	ES Top-500	MS-DB-v4-GPL-CS
3	query	ES Top-1000	MS-DB-tas-b-GPL-all
4	query, topic	ES Top-100	MS-DB-tas-b-GPL-all
5	query, topic	ES Top-500	MS-DB-tas-b-GPL-all
6	query, topic	ES Top-1000	MS-DB-tas-b-GPL-all
7	query	ES Top-500	MS-DB-v4-GPL-CS
8	gen query	ES Top-100	MS-DB-tas-b-GPL-all
9	topic	ES Top-500	MS-DB-tas-b-GPL-all
10	gen topic	ES Top-100	MS-DB-tas-b-GPL-all

- Selection based on 2023 performance
- ElasticSearch API used for creating corpus
- Re-ranked with fine-tuned models

02 TASK 1: Results

Table: Performance of Official Runs on the 2024 SimpleText Task 1 Train Qrels

Run	P@10	R@10	RR@10	nDCG@5	nDCG@10	nDCG@50	nDCG@100
1	**0.612**	**0.103**	**0.799**	**0.584**	**0.555**	0.399	**0.407**
2	0.584	0.088	0.727	0.566	0.550	**0.401**	0.364
3	0.552	0.091	0.761	0.547	0.511	0.369	0.352
4	0.500	0.076	0.666	0.487	0.468	0.356	0.330
5	0.508	0.079	0.657	0.500	0.461	0.353	0.335
6	0.472	0.072	0.697	0.471	0.439	0.337	0.327
7	0.344	0.044	0.470	0.373	0.340	0.227	0.210
8	0.340	0.042	0.502	0.328	0.321	0.236	0.227
9	0.312	0.040	0.451	0.324	0.298	0.205	0.191
10	0.244	0.026	0.309	0.253	0.234	0.160	0.138

- Best: “Query”, "query, topic”
- Worst: generated and topic-level queries
- ↑ corpus size ↓ performance

02 TASK 1: Experiments

Run	Query Input	Corpus	Model
1	query	ES Top-100	MS-DB-tas-b-GPL-all
2	query, topic	ES Top-500	MS-DB-v4-GPL-CS
3	query	ES Top-1000	MS-DB-tas-b-GPL-all
4	query, topic	ES Top-100	MS-DB-tas-b-GPL-all
5	query, topic	ES Top-500	MS-DB-tas-b-GPL-all
6	query, topic	ES Top-1000	MS-DB-tas-b-GPL-all
7	query	ES Top-500	MS-DB-v4-GPL-CS
8	gen query	ES Top-100	MS-DB-tas-b-GPL-all
9	topic	ES Top-500	MS-DB-tas-b-GPL-all
10	gen topic	ES Top-100	MS-DB-tas-b-GPL-all

02

TASK 1: Results

runid	MRR	P@10	P@20	NDCG@10	NDCG@20	Bpref	MAP
AIIRLab_Task1_LLaMABiEncoder	0.9444	0.8167	0.5517	0.6170	0.5166	0.3559	0.2304
AIIRLab_Task1_LLaMAReranker2	0.9300	0.7933	0.5417	0.5943	0.5004	0.3495	0.2177
AIIRLab_Task1_LLaMAReranker	0.8944	0.7967	0.5583	0.5889	0.5011	0.3541	0.2200
...							
UBO_Task1_TFIDFT5	0.7132	0.4833	0.3817	0.3474	0.3197	0.2354	0.1274
LIA_bool	0.7242	0.5233	0.3633	0.3381	0.2891	0.2661	0.1199
Elsevier@SimpleText_task_1_run8	0.7123	0.4533	0.3367	0.3146	0.2752	0.1582	0.0906
Elsevier@SimpleText_task_1_run4	0.6162	0.4300	0.3217	0.3063	0.2681	0.1642	0.1005
Elsevier@SimpleText_task_1_run10	0.5117	0.4067	0.2767	0.2885	0.2365	0.1236	0.0729
AB_DPV_SimpleText_task1_results_FKGL	0.6173	0.3733	0.2900	0.2818	0.2442	0.1966	0.1078
LIA_elastic	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325
Ruby_Task_1	0.5470	0.4233	0.3533	0.2756	0.2671	0.1980	0.1110
LIA_meili	0.6386	0.4700	0.2867	0.2736	0.2242	0.2377	0.0833
Elsevier@SimpleText_task_1_run6	0.5333	0.3833	0.3117	0.2633	0.2430	0.1841	0.0973
Tomislav_Rowan_SimpleText_T1_2	0.5444	0.3733	0.2750	0.2443	0.2183	0.0963	0.0601
Elsevier@SimpleText_task_1_run5	0.4867	0.3533	0.2883	0.2408	0.2232	0.1834	0.0943
Elsevier@SimpleText_task_1_run1	0.5589	0.3000	0.3300	0.2247	0.2399	0.1978	0.1018
Elsevier@SimpleText_task_1_run7	0.4026	0.3200	0.2250	0.2168	0.1850	0.1085	0.0565
Elsevier@SimpleText_task_1_run9	0.3868	0.3300	0.2283	0.2105	0.1829	0.1103	0.0590
Elsevier@SimpleText_task_1_run3	0.4733	0.2367	0.2033	0.1853	0.1703	0.1587	0.0714
Elsevier@SimpleText_task_1_run2	0.4193	0.2233	0.2433	0.1803	0.1865	0.1768	0.0820
Sheringans_Task1_marco_GPT3	0.6667	0.0667	0.0333	0.1149	0.0797	0.0107	0.0107

- Discrepancy between Train qrels and Test qrels rankings
- Best: Gen query, query, topic, gen topic
- Worst: Query and topic
- ↑ corpus size ↓ performance

03

TASK 3: *Rewrite this!*

Simplify passages from scientific abstracts given a query

03

TASK 3

- Simple zero-shot prompting
- Zero-shot prompting with detailed instructions
- Few-shot prompting → in-context learning w.o. updating model params
  - Provided train set as input-output examples.
  - Randomly selected samples.
- Adding Background Information → identify essential information
  - Method 1: *sentence-level simplification* → avoid over-simplification

    Include abstract as context to the sentence
  - Method 2: *breaking down complex terms* → aid lexical simplification

    1. Identify key concepts in given abstract
    2. Provide definitions or synonyms for these concepts

03

TASK 3: Experiments

Table:Configurations of official submissions for Task 3

Run	Prompt	Few-Shot	Level	Two-Step	Uses Abstract
1	1	False	Sentence	False	False
2	2	False	Abstract	False	-
3	3	False	Sentence	False	False
4	4	False	Sentence	False	False
5	2	True	Abstract	False	-
6	5	False	Sentence	True	True
7	6	False	Sentence	False	True
8	7	True	Sentence	False	True
9	8	True	Sentence	False	True
10	6	True	Sentence	False	True
11	8	False	Sentence	False	True
12	5	True	Sentence	True	True

03

TASK 3: Results

Table: Performance of Official Runs on the 2024 SimpleText Task 3 Test Set

FKGL	BLEU	SARI	Run	Prompt	Few-Shot	Level	Two-Step	Uses Abstract
11.54	0.15	36.63	**1**	1	False	Sentence	False	False
12.12	0.12	34.92	**2**	2	False	Abstract	False	-
13.09	**0.25**	**42.57**	**3**	3	False	Sentence	False	False
12.85	0.20	39.00	**4**	4	False	Sentence	False	False
13.26	0.14	36.39	**5**	2	True	Abstract	False	-
13.70	0.21	39.95	**6**	5	False	Sentence	True	True
13.80	0.20	39.31	**7**	6	False	Sentence	False	True
13.74	0.20	39.16	**8**	7	True	Sentence	False	True
13.68	0.21	39.12	**9**	8	True	Sentence	False	True
13.82	0.20	39.05	**10**	6	True	Sentence	False	True
13.70	0.20	38.92	**11**	8	False	Sentence	False	True
13.97	0.19	38.54	**12**	5	True	Sentence	True	True

↑ BLEU, SARI, FKGL
higher similarity, and higher education level

↓ FKGL, BLEU, SARI
simpler, but less similar to reference

Test Set has high FKGL
Sentence-Level: 13.62
Abstract-Level: 13.82

03 TASK 3.1: Results

Table: Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU
References	578	8.86	100	100
Identity	578	13.65	12.02	19.76
Elsevier@SimpleText_Task3.1_run1	578	10.33	43.63	10.68
Elsevier@SimpleText_Task3.1_run4	577	11.73	43.14	12.08
Elsevier@SimpleText_Task3.1_run8	577	12.40	42.95	12.35
Elsevier@SimpleText_Task3.1_run6	577	12.65	42.88	11.76
Elsevier@SimpleText_Task3.1_run7	577	12.55	42.87	12.20
Elsevier@SimpleText_Task3.1_run9	577	12.53	42.61	12.15
Elsevier@SimpleText_Task3.1_run3	577	11.50	42.58	15.75
Elsevier@SimpleText_Task3.1_run10	577	12.57	42.49	11.91
AIIRLab_Task3.1_llama-3-8b_run1	578	8.39	40.58	7.53
AIIRLab_Task3.1_llama-3-8b_run3	578	9.47	40.36	6.26
AIIRLab_Task3.1_llama-3-8b_run2	578	10.33	39.76	5.46
UZH_Pandas_Task3.1_simple_with_cot	578	13.74	39.59	3.38

Discrepancy between test set and official results

FKGL: 13.62 vs 8.86

03

TASK 3.1: Results

Table: Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU
References	578	8.86	100	100
Identity	578	13.65	12.02	19.76
Elsevier@SimpleText_Task3.1_run1	578	10.33	43.63	10.68
Elsevier@SimpleText_Task3.1_run4	577	11.73	43.14	12.08
Elsevier@SimpleText_Task3.1_run8	577	12.40	42.95	12.35
Elsevier@SimpleText_Task3.1_run6	577	12.65	42.88	11.76
Elsevier@SimpleText_Task3.1_run7	577	12.55	42.87	12.20
Elsevier@SimpleText_Task3.1_run9	577	12.53	42.61	12.15
Elsevier@SimpleText_Task3.1_run3	577	11.50	42.58	15.75
Elsevier@SimpleText_Task3.1_run10	577	12.57	42.49	11.91
AIIRLab_Task3.1_llama-3-8b_run1	578	8.39	40.58	7.53
AIIRLab_Task3.1_llama-3-8b_run3	578	9.47	40.36	6.26
AIIRLab_Task3.1_llama-3-8b_run2	578	10.33	39.76	5.46
UZH_Pandas_Task3.1_simple_with_cot	578	13.74	39.59	3.38

Simplest prompts are best at *simplification*

Lowest FKGL

03

TASK 3.1: Results

Table: Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU
References	578	8.86	100	100
Identity	578	13.65	12.02	19.76
Elsevier@SimpleText_Task3.1_run1	578	10.33	43.63	10.68
Elsevier@SimpleText_Task3.1_run4	577	11.73	43.14	12.08
Elsevier@SimpleText_Task3.1_run8	577	12.40	42.95	12.35
Elsevier@SimpleText_Task3.1_run6	577	12.65	42.88	11.76
Elsevier@SimpleText_Task3.1_run7	577	12.55	42.87	12.20
Elsevier@SimpleText_Task3.1_run9	577	12.53	42.61	12.15
Elsevier@SimpleText_Task3.1_run3	577	11.50	42.58	15.75
Elsevier@SimpleText_Task3.1_run10	577	12.57	42.49	11.91
AIIRLab_Task3.1_llama-3-8b_run1	578	8.39	40.58	7.53
AIIRLab_Task3.1_llama-3-8b_run3	578	9.47	40.36	6.26
AIIRLab_Task3.1_llama-3-8b_run2	578	10.33	39.76	5.46
UZH_Pandas_Task3.1_simple_with_cot	578	13.74	39.59	3.38

Zero-Shot vs **Few-Shot**

Examples used in few-shot too complex

03 TASK 3.2: Results

Table: Results for CLEF 2024 SimpleText Task 3.2 abstract-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU
References	103	8.91	100.00	100.00
Identity	103	13.64	12.81	21.36
AIIRLab_Task3.2_llama-3-8b_run1	103	9.07	43.44	11.73
AIIRLab_Task3.2_llama-3-8b_run2	103	10.22	42.19	7.99
AIIRLab_Task3.2_llama-3-8b_run3	103	10.17	43.21	11.03
Elsevier@SimpleText_Task3.2_run2	103	11.01	42.47	10.54
Elsevier@SimpleText_Task3.2_run5	103	12.08	42.15	10.96

Zero-Shot vs Few-Shot

Examples used in few-shot too complex

03

TASK 3: Results

```
### TASK ###
Simplify the language used in this sentence from a scientific article so that it can be understood by the general audience.
Focus on simplifying the sentence structure and replacing scientific jargon with everyday language.

### REQUEST ###
- Sentence:
{row.source_snt}

- Simplified Sentence:
```

04

Conclusion

Task 1

- Performance not as high as the previous year
- Hypothesis: shift from lexical to semantic search models and generative methods in the reference set
- Generated Search Queries > traditional search queries
- Best performance with query-level generated queries

Task 3

- Various prompt-engineering techniques tested
- Simplest prompts yielded best FKGL and BLEU scores
- Few-shot prompts underperformed due to complexity mismatch between test and reference sets

CLEF 2024
GRENOBLE

SIMPLE TEXT

09-09-2024

Enhancing Scientific Document Simplification through Adaptive Retrieval and Generative Models

Artemis Capari, Hosein Azarbonyad, Zubair Afzal, Georgios Tsatsaronis

ELSEVIER