SimpleText Pilot Task Guidelines

We invite you to submit both automatic and manual runs! Manual intervention should be reported.

Access

Please register at the SimpleText@CLEF workshop in order to access the data:
http://clef2021-labs-registration.dei.unipd.it/
After registration, you will receive an email with information on how to log in to the data server:
https://guacamole.univ-avignon.fr

Pilot Task 1: Selecting passages to include in a simplified summary - Content Simplification

Given an article from a major international newspaper general audience, this pilot task aims at retrieving from a large scientific bibliographic database with abstracts, all passages that would be relevant to illustrate this article. Extracted passages should be adequate to be inserted as plain citations in the original paper. .

2021 DataSet

For this edition we use the Citation Network Dataset: DBLP+Citation, ACM Citation network (https://www.aminer.org/citation). An elastic search index is provided to participants accessible through a GUI API. This Index is adequate to:

- apply basic passage retrieval methods based on vector or language IR models
- generate Latent Dirichlet Allocation models,
- train Graph Neural Networks for citation recommendation as carried out in https://stellargraph.readthedocs.io/ for example,
- apply deep bi directionnal transformers for query expansion.
- and much more …

2021 Queries

- For this edition queries are a selection of recent n press titles from The Guardian enriched with keywords manually extracted from the content of the article. It has been checked that each keyword allows to extract at least 5 relevant abstracts. The use of these keywords is optional.

Format

Input:

- Topics are in the following format:

```
<topic>
    <topic_id>topic_id</topic_id>
    <topic_text>topic_text</topic_text>
    <keywords>
        <keyword>keyword_1</keyword>
        <keyword>keyword_2</keyword>
        ...
        <keyword>keyword_n</keyword>
    </keywords>
</topic>
```

ElasticSearch index on the following data server (participant login required):
https://guacamole.univ-avignon.fr/dblp1/_search

Output:

A maximum of 1000 passages to be included in a simplified summary in a TSV (Tab-Separated Values) file with the following fields:

- *run_id:* Run ID starting with *team_id_*
- *manual:* Whether the run is manual {0,1}
- *topic_id:* Topic ID
- *doc_id:* Source document ID
- *passage:* Text of the selected passage
- *rank:* Passage rank

run_id *manual* *topic_id* *doc_id* *passage* *rank*

Evaluation

Sentence pooling and automatic metrics will be used to evaluate these results. The relevance of the source document will be evaluated as well as potential unresolved anaphora issues.

Example:

INPUT

```
<topic>
```

```
        <topic_id>1</topic_id>
        <topic_text>Digital assistants like Siri and Alexa entrench gender biases, says
        UN</topic_text>
        <keywords>
                <keyword>Digital assistant
                </keyword>
                <keyword>Biases</keyword>
        </keywords>
</topic>
```

OUTPUT

run_id	manual	topic_id	doc_id	passage	rank
ST_1	1	1	3000234933	People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects.	1
ST_1	1	1	3003409254	big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc.	2
ST_1	1	1	3003409254	Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.	3

Pilot Task 2: Identifying difficult-to-understand concepts for non experts - Content Simplification

The goal of this pilot task is to decide which terms (up to 10) require explanation and contextualization to help a reader to understand a complex scientific text - for example, with regard to a query, terms that need to be contextualized (with a definition, example and/or use-case).

Format

Input:

- Topics in the following format:

```
<topic>
    <topic_id>topic_id</topic_id>
    <topic_text>topic_text</topic_text>
    <passage_id>passage_id</passage_id>
    <passage_text>passage_text</passage_text>
</topic>
```

Output:

List of terms to be contextualized in a tabulated file TSV with the following fields:

- *run_id:* Run ID starting with *team_id_*
- *manual:* Whether the run is manual {0,1}
- *topic_id:* Topic ID
- *passage_id:* Passage id
- *term:* Term or other phrase to be explained
- *rank:* Importance of the explanation for a given term

Run_id manual topic_id passage_id term rank

Evaluation

Term pooling and automatic metrics (NDCG,...) will be used to evaluate these results.

Example:

Input:

```
<topic>
    <topic_id>1</topic_id>
    <topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>
    <passage_id>1</passage_id>
    <passage_text>Automated decision making based on big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination. The majority of thus far
```

```
        proposed fairness-aware machine learning approaches focus solely on the
        pre-, in- or post-processing steps of the machine learning process, that
        is, input data, learning algorithms or derived models, respectively.
        However, the fairness problem cannot be isolated to a single step of the
        ML process. Rather, discrimination is often a result of complex
        interactions between big data and algorithms, and therefore, a more
        holistic approach is required. The proposed FAE (Fairness-Aware
        Ensemble) framework combines fairness-related interventions at both pre-
        and postprocessing steps of the data analysis process. In the
        preprocessing step, we tackle the problems of under-representation of
        the protected group (group imbalance) and of class-imbalance by
        generating balanced training samples. In the post-processing step, we
        tackle the problem of class overlapping by shifting the decision
        boundary in the direction of fairness. </passage_text>
    </topic>
```

Output:

run_id	manual	topic_id	passage_id	Term	rank
ST_1	1	1	1	machine learning	1
ST_1	1	1	1	societal biases	2
ST_1	1	1	1	ML	3

Pilot Task 3: Scientific text simplification - Language simplification

The goal of this pilot task is to provide a simplified version of text passages. Participants will be provided with queries and the passages from the abstracts of scientific papers.

Format

Input:

- Topics in the following format:

```
    <topic>
        <topic_id>topic_id</topic_id>
        <topic_text>topic_text</topic_text>
        <passage_id>passage_id</passage_id>
        <passage_text>passage_text</passage_text>
```

</topic>

Output:

Simplified passages in a TSV tabulated file with the following fields:

- *run_id:* Run ID starting with *team_id_*
- *manual:* Whether the run is manual {0,1}
- *topic_id:* Topic ID
- *passage_id:* Passage id
- *simplified_passage:* Text of the simplified passage

Run_id manual topic_id passage_id simplified_passage

Evaluation

The simplified passages will be evaluated manually with eventual use of aggregating metrics.

Example

Input:

<topic>
<topic_id>1</topic_id>
<topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>
<passage_id>1</passage_id>
<passage_text>`Automated decision making based on big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.` *</passage_text>*
</topic>

Output*:*
Run_id manual topic_id passage_id simplified_passage
ST_1 1 1 1 Automated decision-making may include sexist and racist biases because their algorithms are based on the most prominent social representation in the dataset they use.