Summarization

The summarization module uses relevance judgments from CLIR to generate corresponding summaries for each relevant query-document pair that shows the most relevant translation for each query.

Input

CLIR output: a json config file, UMD-WorkECDir, UMD-WorkECDir-f1 (optional, for retranslation)
Data directory: Source documents, queries, and translations as indicated by data_structure file, query_processor directory

Output

(please also indicate if this component modifies the input in-place)
The inputs are not modified in place

Output directory consisting of the following folders: annotations, images, markup, retranslation, packages.

Docker Commands

```
docker run \
    -v <parameter A> \
    -v <parameter B> \
    -v <parameter C> \
    -v <parameter D> \
    -v <parameter E> \
    <parameter F> \
    <parameter G> \
    <parameter H> \
    <parameter I> \
    <parameter J>
    ...
```

- <parameter A>: $NIST_VOL:/NIST-data
- <parameter B>: $EXPERIMENT_VOL:/experiment
- <parameter C>: $CLIR_VOL:/clir
- <parameter D>: $OUTPUT_DIR:/outputs
- <parameter E>: /var/run/docker.sock:/var/run/docker.sock
- <parameter F>: docker name
- <parameter G>: $RUN_NAME
- <parameter H>: $OUTPUT_DIR
- <parameter I>: $NUM_PROCS

- <parameter J>: $GPU_ID (optional if retranlation is not used (i.e. not kk or ka))

Examples

```
NIST_VOL=/storage/data/NIST-data
EXPERIMENT_VOL=/storage/proj/dw2735/experiments/docker_test/ps/text
OUTPUT_DIR=/storage/proj/dw2735/summarizer_output/docker_test/ps/text
CLIR_VOL="$EXPERIMENT_VOL/UMD-CLIR-workECDir"

RUN_NAME="CUSUM"
NUM_PROCS=12
GPU_ID=0

docker run -it --rm \
        --user "$(id -u):$(id -g)" \
        --group-add $(stat -c '%g' /var/run/docker.sock) \
        --ipc=host \
        --name summarizer \
        -v $NIST_VOL:/NIST-data \
        -v $EXPERIMENT_VOL:/experiment \
        -v $CLIR_VOL:/clir \
        -v $OUTPUT_DIR:/outputs \
        -v /var/run/docker.sock:/var/run/docker.sock \
        summarizer:v3.0 \
        $RUN_NAME \
        $OUTPUT_DIR \
        $NUM_PROCS \
        $GPU_ID
```

System Requirements

(list only the ones that are applicable)

- CPU: 1 (x86_64)
- RAM: 32GB
- GPU: same requirement as scriptsmt. Required for retranslation only.
- GPU-RAM: same requirement as scriptsmt. Required for retranslation only.
- Target CUDA version and/or minimum NVIDIA driver version (current Scripts servers have CUDA: 11.2 and Driver: 460.32.03): same requirement as scriptsmt. Required by retranslation only.

Standalone

No. For retranslation, docker requires scriptsmt or ape.

Approach

We use various annotators (query relevance predictor, psq, lexical match, stem matches) to score sentences, and use Borda Count to rank and select the top ranking sentences up to the word budget. For Kazakh and Georgian, we also provide retranslation, which takes in generated summaries and retranslates some of the sentences to include the query word.

Notes

Code for docker available: https://github.com/eturcan/scripts