Tokyo Tech Student Summarization with Gaze Information (SSG23)

Synopsis

SSG23 is a student summary corpus with eye-tracking information from two English source texts: Cycloclean (cyc) and Napping and Learning (nap). The students were asked to read the source texts and summarise their main ideas and key details in approximately 80 words in English. The dataset comprises 53 XML files output from Translog-II. The summaries were written by 30 Japanese university students (undergraduate and graduate level). They are non-native speakers of English.

Corpus collection

The corpus collection experiment captures both implicit events from eye gaze and explicit events of marking important text spans during summarisation. The data was collected by Translog-II. We modified the program to implement a feature highlighting text spans in the source text. The highlighting events of text spans are recorded in the event log data along with other event data, such as fixation and keystroke events. The participants worked on a 23.8-inch LCD monitor with an infrared eye-tracker Tobii Pro X3-120, which has a sampling rate of 120Hz. To prevent intervention from OS, e.g. popup menus, the keyboard keys other than alphanumeric, symbol, return and delete keys were disabled. Before starting the experiments, the participants must watch a 25-minute lecture video explaining the summarisation. Before the experiment, the participants were explained the interface with a sample English text. The experiments started by calibrating the eye tracker with each participant. After the calibration, the participants were instructed to do a full reading first and highlight important information, then continue to write the summary while having access to the source texts and their highlighted information. The participants were encouraged to write an abstractive summary of about 80 words in English. The participants were fairly compensated for their participation, with some monetary benefits of 3,000 JPY for 90 minutes.

File formats

The Translog-II XML output contains:

- Fixation event from Tobii X-120 eye-tracker device
- Keyboard events of character insertion and deletion on summary fields
- Mouse events to underline text spans

File Format: Text: XML, UTF-8

Directories

```
SSG23
├─ mapping.txt
├─ preprocessed
|    ├─ Alignment
|    ├─ Events
|    ├─ Tables
|    └─ Translog-II
├─ raw
├─ README.md
├─ source_text
└─ summary
```

Raw File name structure

File name: TT{Subject_ID}-{text_name}.xml Attributes:

- Subject_ID: the unique identifier of participant ID
- text_name: the title of the source text to be summarized: "cyc" or "nap".

Example: TT01-cyc.xml

Preprocessing

There is a preprocessing tool that converts Translog-II log files into convenient formats for analysis. To process the data, follow the TPRDB Perl script instruction:

1. Rename the XML logging file to the P{ID}_{TaskTextID}.xml (File name mapping is found in "mapping.txt".)
2. Tokenize texts `./StudyAnalysis.pl -T tokenize -D SSG2023`
3. Generate table files `./StudyAnalysis.pl -C tables -S SSG2023`

After running the above three steps, the following directories are created, which include the converted files.

- **Alignment**: This directory contains the output of tokenized XML text into three different suffixes: ''.*src,* ''.tgt and ''*''.atag.
- **Events**: This directory contains each XML file filtered keyboard, mouse and eye-tracking events into Atag.xml and Events.xml.
- **Tables**: This directory contains tab-separated files with different suffixes. Each file contains some features. In total, there are almost 400 features that can be used to describe and model behaviour during translation. Further details tpr-db/features.
- **Translog-II**: This directory contains raw XML files named following the TPRDB Perl script convention

Acknowledgement

This data construction was supported by JSPS KAKENHI Grant Number JP20H01292 (PI: Prof. Sawaki, Yasuyo at Waseda University). The author of the source texts: Cycloclean and Napping & Learning is Sawaki, Yasuyo (ysawaki@waseda.jp).