SEGMENTANNO: An annotation tool for text segmentation*

Viet Dac Lai[1,*], Franck Dernoncourt[2] and Thien Huu Nguyen[1]

[1]Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
[2]Adobe Research, Seattle, WA, USA

Abstract

We present Segmentanno, a fast light-weight interactive general-purpose annotation tool for text segmentation. Segmentanno has been specifically designed to speed up the annotation for text segmentation by minimizing mouse activities. Segmentanno uses an interactive web interface for making text boundaries and labeling text units. Segmentanno can be used for various levels of text units (e.g. token, clause, sentence, and paragraph) and various tasks such as sequence labeling, text classification. Segmentanno is available with an open-source license at https://github.com/laiviet/segmentanno

Keywords

Annotation Tool, Text Segmentation, Video Transcripts

1. Introduction

Generating large-scale high-quality corpora is crucial to the development of natural language processing (NLP). However, achieving both large-scale and high-quality targets are extremely time-consuming and expensive in many NLP tasks. Even though some NLP tasks requires highly professional linguists to produce the finest annotated data [1], many other NLP tasks can benefits from remote crowd-sourcing annotation [2] to achieve numerous annotations in a short period of time. However, the bottle neck lies in the level of skillfulness of the annotators and the employed annotation tool. There have been many general purpose annotation tools developed for crowdsourcing such as BRAT [3], WebAnno [4], and Doccano [5]. Their broad coverage leads to an adverse higher complexity of the human-machine interface, which results in more human actions needed to perform the annotation.

Data annotation actions for NLP tasks are focused. They includes typing some words, selecting spans, assigning a label to a text unit, linking two text units. Among these actions, label assigning and unit linking is currently done very efficiently by using keyboard shortcut and mouse dragging. In contrast, span selection and word typing are the most time consuming, hence, there are rooms for improvement. Therefore, to improve annotation productivity and accuracy, annotation tools should reduce the complexity of the span selection action. An

VTU'22: The AAAI-2022 Workshop On Video Transcript Understanding, Feb 28, 2022, Virtual

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

vietl@cs.uoregon.edu (V. D. Lai); franck.dernoncourt@adobe.com (F. Dernoncourt); thien@cs.uoregon.edu (T. H. Nguyen)

example is text segmentation, i.e. split text into paragraph by adding a boundary between sentences. In fact, this process can be seen in other similar tasks such as subtitle segmentation and punctuation restoration. In subtitle segmentation, the subtitle is split into meaningful chunk with a limit of number of characters. In punctuation restoration, one adding punctuation marks such as comma, period, question mark to split sentences, clauses and phrases. In this setting, assigning a label, aka. a punctuation marker, to a token means the token is followed by the the marker. Translating this task to set of annotation actions, an annotator need to choose a token, then, assign a label to it. The span selection in this tasks is virtually unnecessary as the task is on the token level only. We argue that for this set of tasks, a dedicated annotation tool can expedite the training of annotator and the annotation process.

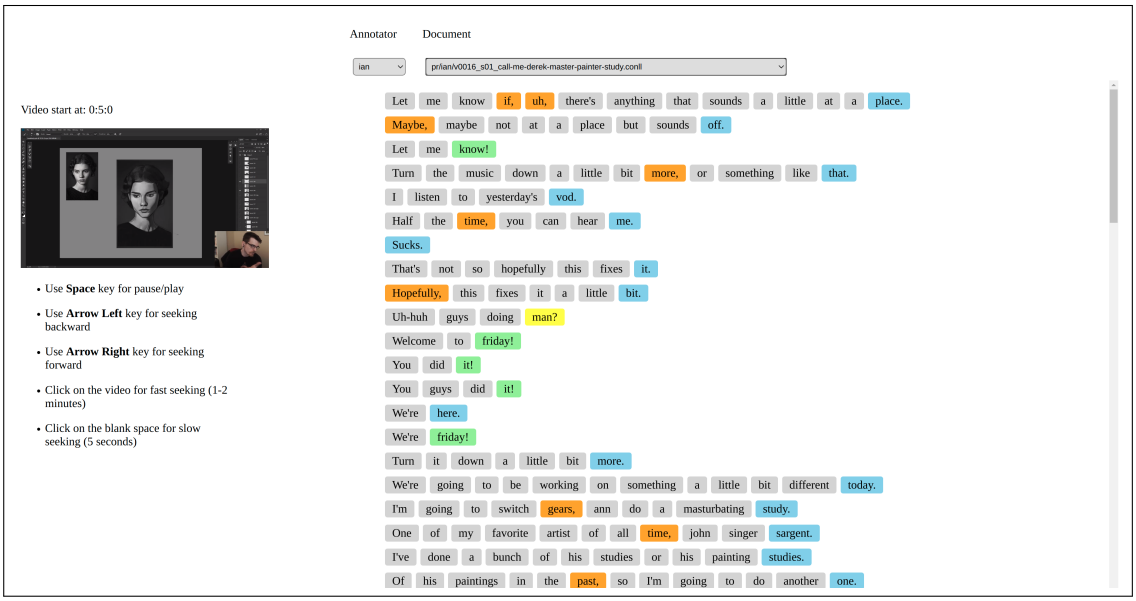

Figure 1: Visualization of Segmentanno interface

In this paper, we introduce the Segmentanno (SEGMENTation ANNOtation), an open-source annotation tool designed for text segmentation. Segmentanno provides all annotation functionalities through web-based interface making it highly convenient for crowd-sourcing environment. The interface can be easily customized using web standard and NLP community format to specific tasks. In the mean time, its interactive interface ensure that the annotated texts are human-readable friendly. This feature can create high quality annotation as it facilitate the annotator to contextualize the text content.

In section 2, we present its functionalities and its general usage. In section 3, we present two case studies of using the Segmentanno in annotating a corpus for punctuation restoration with 500,000 tokens and a chitchat detection of xxx sentences. We also conduct an human survey to compare human performances on different annotation tools.

2. Annotation with Segmentanno

Segmentanno is a web-based easy-to-use designed for text segmentation, which focuses on the productivity and experiences of all related parties: annotators, curators, and organizers. In Segmentanno, the annotation is organized into three levels: project, annotator, and document.

2.1. Interface

The user interface of the website is designed to focus on easy usage and productivity by reducing the mouse movement. It is developed to be used by common browsers such as Chrome-family, Safari, and Firefox.

Once accessing the website, the annotator first need to select the annotator from the list, then, the list of assigned documents will be automatic loaded. The annotator then choose the document he/she want to work with. The document will be loaded and presented on the right side of the interface. The text is presented as a sequence of rectangular box with colorful background. We propose two changes to the visualization compared to previous tools that significantly improve the readability of the annotated text. First, instead of presenting the label ontop of the box like BRAT and Doccano, Segmentanno insert the marker directly into the box. Second, if the marker is a termination of a sentence or a paragraph, it will automatically create a new line. Similar to the other tool, Segmentanno allow text color and background color configuration to make the visualization more vivid.

If the video is available, a video playback is presented on the left side of the interface. The video playback will automatically load the video, seek the video cursor to the starting points of the document, and automatically play the video.

2.2. Labeling

With the target to boost the productivity, mouse click and mouse movement are reduced to the minimum. First, to select the span, the annotator just need to hover over the text unit. Then, a popup will appear below the token to show the label option. This significantly reduce the burden of selecting the span on span-based annotation tool like BRAT [3]. And most importantly, it completely eliminates the span error while using mouse to highlight the span. It also reduces the stress on annotator fingers if they have to double clicked to select the word like Doccano [5]. Then one can use click or keyboard shortcut to select the label. If the label is a sentence termination such as period, question mark and exclamation mark. The rest of the text is moved to the next line. And interestingly, if they use the mouse in this case, the move will be put on the next sentence, ready for annotating. We found that this feature is extremely useful for annotating text classification task, e.g. chitchat detection dataset.

Segmentanno supports video/audio playback for transcribing task. As the livestreaming might be very long, e.g. several hours, seeking few seconds using mouse is unworkable. Segmentanno allows the annotator to navigate the video using navigation keys. Each keypress seeks the cursor 5 to 10 seconds.

2.3. Data curation

Once the annotation phase is completed, a data curation phase follows to merge the annotation of one annotator into the one of the other annotator. The annotator can view the annotation of both and fix the disagreement.

2.4. Running an annotation project

Segmentanno is a platform-indepdent annotation tool developed using Python and Django framework. For each project, the organizer of the project creates a project folder with a subfolder for each annotator, similar to the BRAT setting. In each subfolder, the documents are stored in tab-seprated files with .conll extension. If video is needed, another JSON-format with the same name, but .meta.json extension is placed in the same folder. Unlike BRAT, Segmentanno use a single annotation configuration file written in JSON format, named "annotation.json" under the project root. The annotation configuration allows configuration of labels, visualization, and keyboard shortcut.

```
root
└── static
    └── data
        ├── project 1
        │   ├── annotation.json
        │   ├── annotator 1
        │   │   ├── doc 1.conll
        │   │   └── doc 1.meta.json
        │   └── annotator 2
        └── project 2
```

3. Case Study: Punctuation Restoration

Segmentanno has been successfully used to create the Behance-PR punctuation restoration dataset of 180 hours of livestreaming video with 1,300,000 tokens. In particular, we aim to annotate 4 types of punctuation including comma, period, question, and exclamation marks. Figure 2 shows an example of the annotation configuration for BehancePR dataset.

The livestreaming video on Behance has varied lengths ranging from 30 minutes to 5 hours. We split them into shorter clips of approximately 5 minutes so that it can reduce the burden of processing large files and reduce the stress of facing endless text to the annotators. The text is generated by Microsoft Automatic Speech Recognition (ASR) from live-streaming videos. All the transcribed texts are then pre-tokenized and presented in a tab-separated format with the default label "O", and saved as a .conll extension. In order to support video playback, for each document file, the organizer should produce a meta file that contains the information of the video such as URL, and video starting point as presented in figure 3.

```
{
  "label": [
    "COMMA",
    "PERIOD",
    "EXCLAMATION",
    "QUESTION",
    "O"
  ],
  "entity_separator": " ",
  "label_extra": {
    "COMMA": ",",
    "PERIOD": ".",
    "EXCLAMATION": "!",
    "QUESTION": "?",
    "O": ""
  },
  "text_color": {
    "COMMA": "black",
    "PERIOD": "black",
    "EXCLAMATION": "black",
    "QUESTION": "black",
    "O": "black"
  },
  "background_color": {
    "COMMA": "orange",
    "PERIOD": "skyblue",
    "EXCLAMATION": "lightgreen",
    "QUESTION": "yellow",
    "O": "lightgray"
  }
```

Figure 2: Annotation configuration of for the punctuation restoration with Segmentanno.

```
{
  "title": "Livestreaming with Diane Do from Fresco",
  "video_link": "http://streamprod-westus-streamprodwestus-uswe.
      streaming.media.azure.net/279a94ca-634f-473c-b75d-9ab969350048/
      output.mp4",
  "video_start": 6320.58,
  "duration": 278.2399999999998
}
```

Figure 3: Configuration for each document with a title, URL to the video, the starting point of the document in the whole video, and duration of the document.

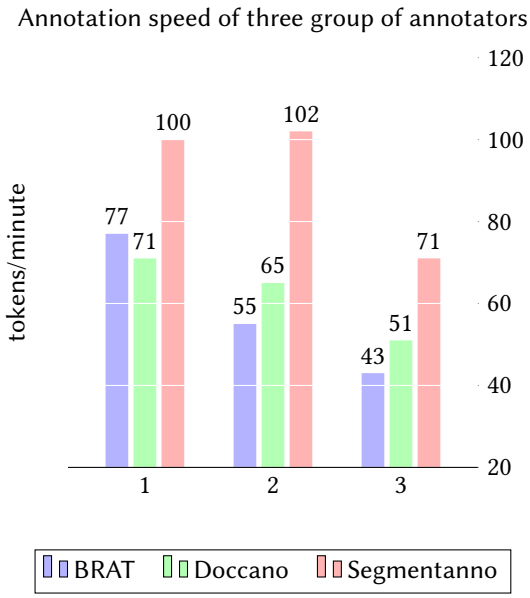

Figure 4: Human evaluation on 3 BRAT, Doccano, and Segmentanno

4. Human evaluation

To evaluate the benefit of using Segmentanno, we hire 3 freelancers. We aim to evaluate human performance on three annotation tools including BRAT, Doccano, and Segmentanno. To isolate the difficulty of the text, we use the same text for three annotators. We also use different tool orders for each of the annotators for a fair comparison. For each tool, a freelancer is given a non-punctuated document of 1000 tokens to do in 20 minutes. The annotator will record the number of tokens and the amount of time needed to do it. The average annotation speed (tokens per minute) is presented in Figure 4.

5. Conclusion

This paper presented Segmentanno, a lightweight open-source web annotation tool designed for text segmentation. Segmentanno targets simple configuration with a community-familiar configuration format. Segmentanno is a platform-independent web application that supports the most common browsers. Segmentanno 's interface is heavily optimized for mouse and shortcut usage which significantly reduces the mouse movements and clicks. The same interface and features can be used for data curation.

Segmentation was used successfully on two datasets including punctuation restoration with 1.3M tokens and chitchat detection with 360K sentences. Segmentanno supports multiple-level text segmentation and text classification e.g. token-level, phrase-level, sentence-level, and paragraph-level. In the future, we plan to improve management functionalities such as GUI-based project creation, dashboard, and agreement management.

References

[1] M. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of english: The penn treebank, Computational Linguistics 19 (1993) 313–330.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[3] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.

[4] R. E. de Castilho, C. Biemann, I. Gurevych, S. M. Yimam, Webanno: a flexible, web-based annotation tool for clarin, in: Proceedings of the CLARIN Annual Conference (CAC), Citeseer, 2014.

[5] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.