International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

NLP Assisted Text Annotation

Authors

[*1] *Chandan S Patil, Department of Computer Engineering and Data Science, Presidency University Bengaluru*

[*2] *G Adarsh Vardhan, Department of Computer Engineering and Data Science, Presidency University Bengaluru*

ABSTRACT

This paper presents the importance and use of text annotation in NLP models and reviews BRAT - an NLP-assisted text annotation tool and other powerful and modern methods.

BRAT is an NLP-based tool for text annotation (annotation: making notes, adding comments to a text in order to give an explanation or opinion to the source article). It is used for high-quality structured annotation for various NLP activities and focuses on increasing annotator productivity using NLP techniques.

Annotation teaches NLP models how to identify parts of speech, real-world objects (named entities such as a person, location, place), and key phrases within a text. NLP algorithms are trained using large, annotated text datasets and annotation is the backbone of services like automatic speech recognition, chatbots, and sentimental analysis.

Keywords: NLP, ML, CNN, TCN, Ensemble, Corpus.

INTRODUCTION

The purpose of the review article is to evaluate the use of Natural Language Processing NLP in text annotation and many tools developed for assisting in such tasks. Many researchers have produced different deep learning models such as BRAT [1], Doccano [2], INCEpTION [3], and GATE - General Architecture for Text Engineering [4]. Moreover, we believe that these models can be used in the fields of Information Retrieval and Organization, Machine Learning, and Artificial Intelligence to enhance the web browsing experience and voice assistants like Google Assistant, and Apple's SIRI. In this review article, we will evaluate and summarize text annotation using NLP and tools such as BRAT and many other tools developed by researchers which are used for efficient and automated text annotation.

BACKGROUND

Manually curated high-quality annotations are the backbone for training and evaluation for most of the Natural Language Processing NLP tasks. Annotation is also one of the most time-taking and cost-intensive elements of many NLP models and research and can place demand for high-quality and consistent human annotators. Modern annotation tools on the other hand are based on technique and offer minuscule support to users beyond the minimum required functionality. They feature Interactive and user-friendly interfaces along with the use of NLP models to assist and not replace human judgment to maintain the quality of annotations and make it accessible to non-technical users. The use of NLP to train annotation tools can reduce both human and financial costs of annotation.

LITERATURE

It is not adequate for a model to be trained using large and diverse data and expect it to accurately detect speech or recognize text, data must be prepared so that the model can find inferences, semantics, overlapping text, and patterns. This is achieved by introducing metadata (annotation) to a dataset. The annotation done should be accurate and relevant to the learning task to gain efficiency and accuracy. Therefore, text annotation plays a significant role in training many machine learning and NLP models and developing intelligent language technologies.

Natural language processing is one of the most rapidly growing areas of Artificial Intelligence research. NLP technologies such as voice assistants, chatbots, and sentimental analysis algorithms have helped enterprises improve their efficiency and production. Recent advances in NLP processing have shown potential to help the speech impaired and people from diverse cultures communicate freely through NLP-based applications. However, without text annotation, none of these remarkable innovations would be developed.

The goal is to minimize manual effort and ensure deduplication (and other downstream data processing) is maximally effective and eliminate the gap between human perception and machine representation.

Deep Learning techniques used for Text Annotation

Most of the models used are TCN (Temporal Convolutional Network) or Ensemble-based modes. TCN is an exceptional alternative to recurrent architecture model RNN since it can take a long sequence of inputs without suffering from vanishing gradients and has been proven to be effective in classifying text data.

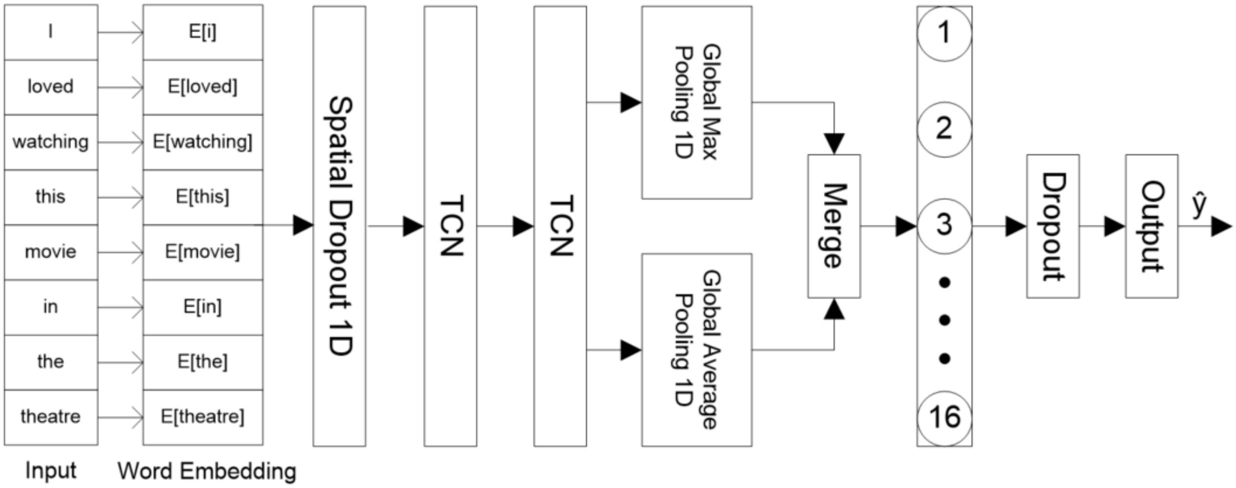


Figure 1. The proposed TCN model.

Ensemble models outperform other single-based models in tasks such as classification and can help make better predictions. Models such as the 1D CNN and BiRNN are excellent models which can be combined for text classification.

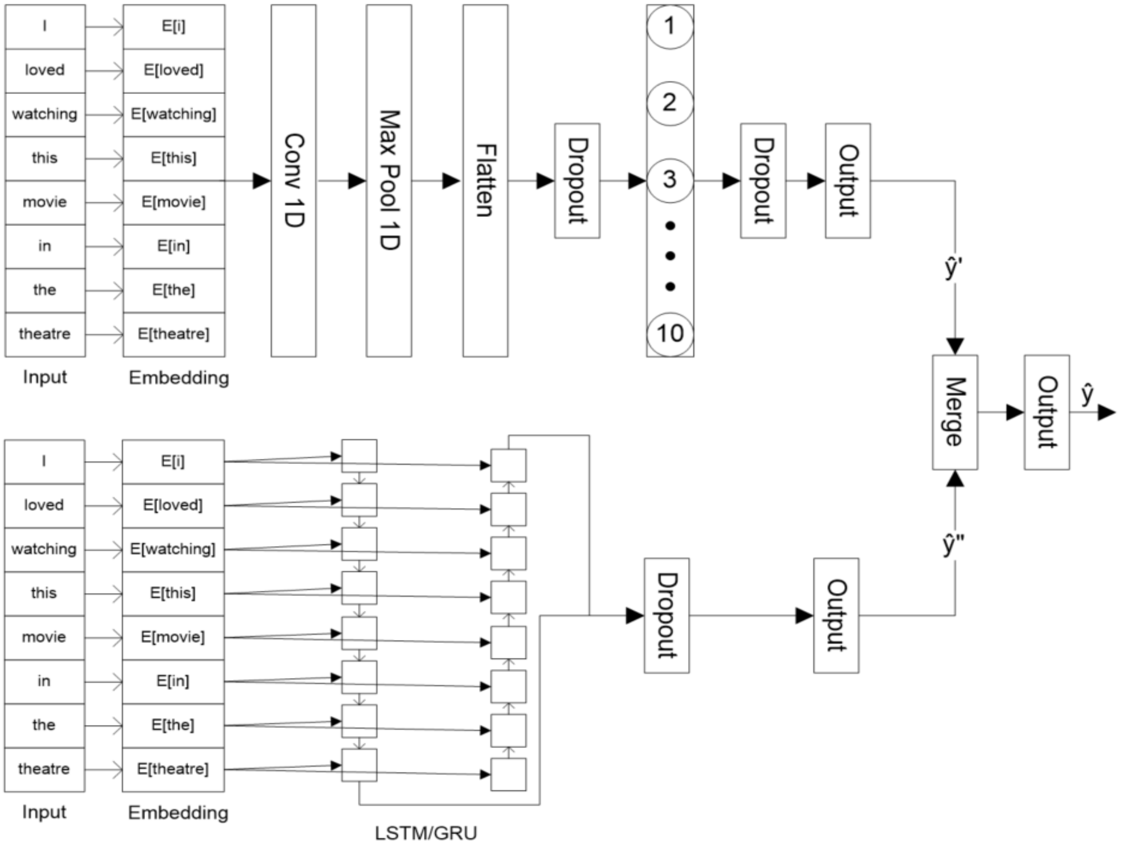


Figure 2. ensemble-based model with CNN and BiGRU.

Studies on various models and their performance concluded that using word embedding caused a spike in performance. Pre-trained models can increase the accuracy of the model to a high margin and using an ensemble-based model with TCN can challenge the benchmarks even more.

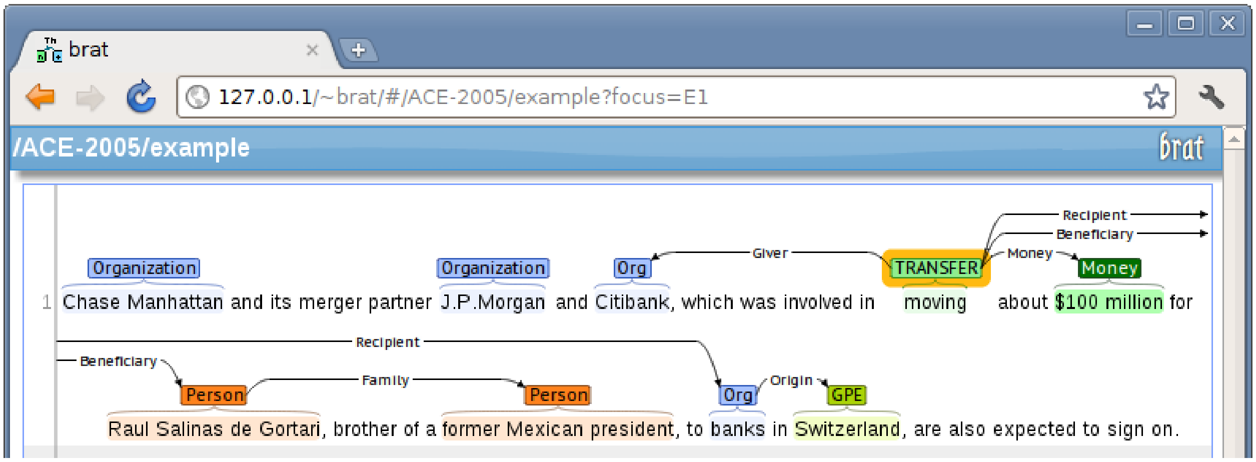
NLP Assisted Annotation Tools

These are algorithms designed for the automated addition of linguistic information to a corpus. They are trained using Deep Neural Networks with large datasets and can be developed to perform several types of annotation such as sentimental annotation, entity annotation, text classification, linguistic annotation, and intent annotation. There are many open-source tools available on the internet and we will be reviewing some of them later in this article.

Features Required for a Good Annotation Tool

High-Quality Annotation Visualization and Interface

The creation of an interactive and user-friendly annotation visualization is critical for users to understand the text being annotated and users from non-tech backgrounds can find it easier to use the tool. Use of a vector-based visualization that provides a scalable and detailed rendering.
Combining PDF and EPS image formats functionality to support use in for example figures in a publication. Features such as annotation editing by using modern web technologies can offer users a familiar environment to work with.



127.0.0.1/~brat/#/ACE-2005/example?focus=E1

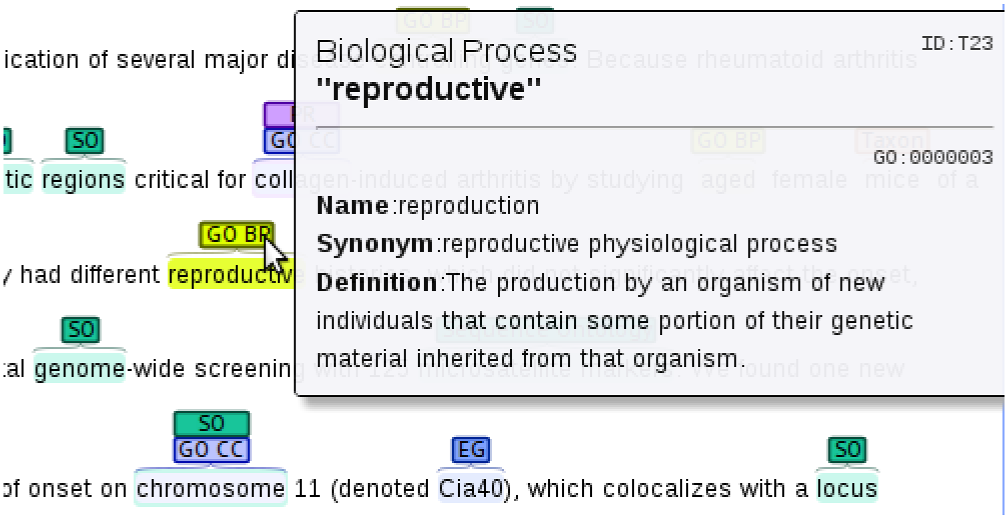
/ACE-2005/example

1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

Figure 3. BRAT UI, connecting annotations for “moving” and “Citibank”

Multifaceted Annotation Support

Annotation tools must be capable of performing diverse annotation tasks such as POS-tagging, normalization, entity mentions, chunking, semantic role labeling, and many more. Such features can enable them to perform tasks such as source detection, text illustration, geolocation identification, and caption generation.



Biological Process ID:T23
"reproductive"
GO:0000003
Name:reproduction
Synonym:reproductive physiological process
Definition:The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism.

ication of several major di
SO
tic regions critical for coll
GO CC
GO BP
/ had different reproductiv
SO
:al genome-wide screenin
SO
GO CC
EG
SO
of onset on chromosome 11 (denoted Cia40), which colocalizes with a locus

Figure 4. BRAT UI, an example of normalization of a corpus (CRAFT)

Integration of NLP technology

Implementation of various NLP technologies such as TCS and Ensemble-based models as mentioned above and various other models such as HMM - Hidden Markov Model can be used to integrate automatic annotation and the annotation workflow. Human judgment cannot be fully replaced. To address this issue, augmentation of input processes from statistical and machine learning models for the annotation while still using a human annotator to supervise the annotation procedure has been seen to be an effective process.

Corpora Search Functionality

Users can search the data using a full range of search tools and the annotated result with a rich set of options using the keyword in context concordance and for browsing using the aspects of the matched annotation (type, text, context).

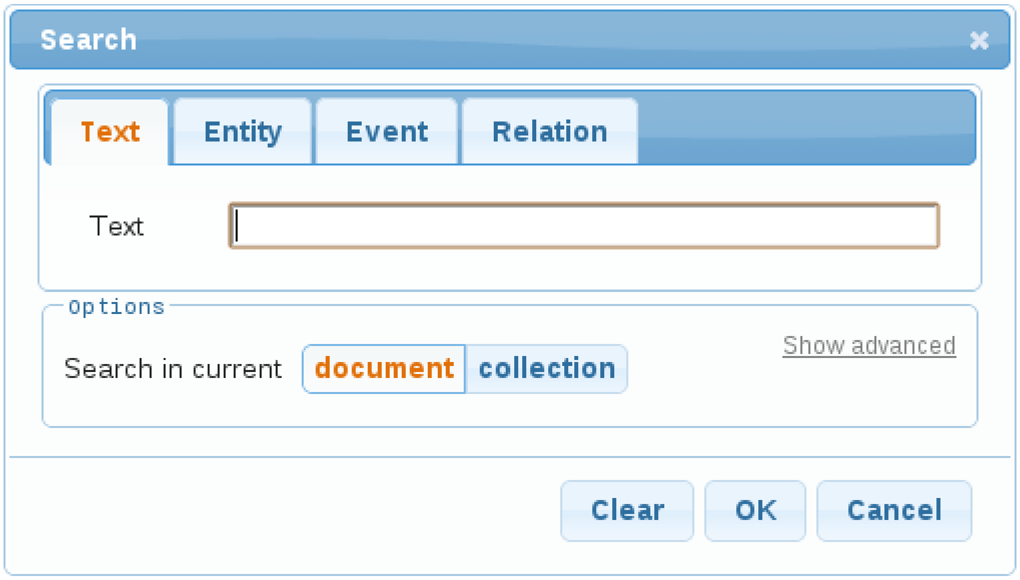
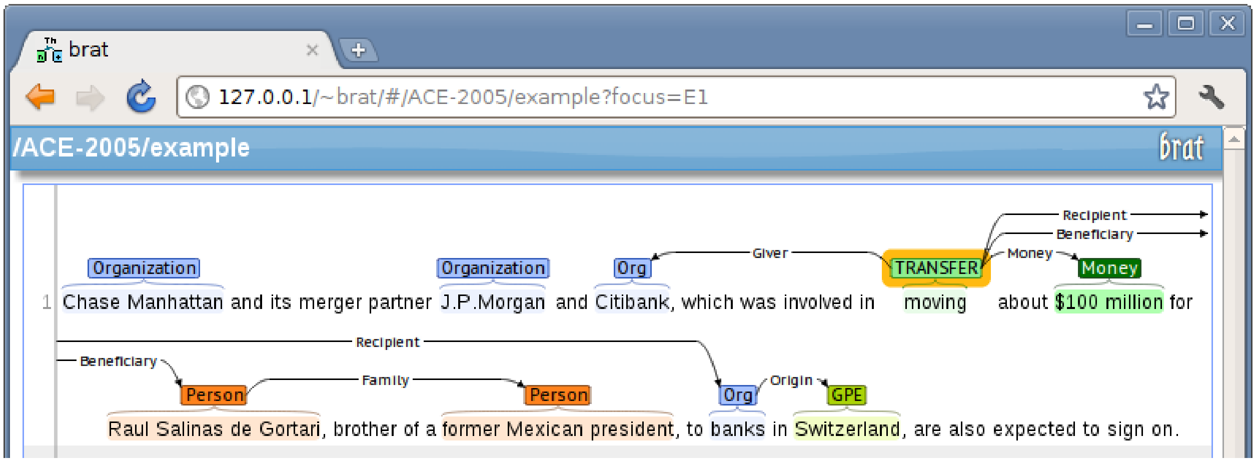


Figure 4. BRAT search dialog

BRAT - NLP Assisted Text Annotation Tool

BRAT is designed for structured annotation, where the text/corpus has a fixed form but is not freeform text that can be processed and interpreted automatically using an ANN-based algorithm.

BRAT also supports the annotation of **n-ary associations**, which can link a large number of annotations that participate in certain roles and is ideal for **named entity recognition** and binary relations for simple **relational information extraction** tasks. It also supports **normalization** annotations that associate other annotations with resources such as Wikipedia. BRAT also implements NLP-based techniques to support human annotation efforts.



1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

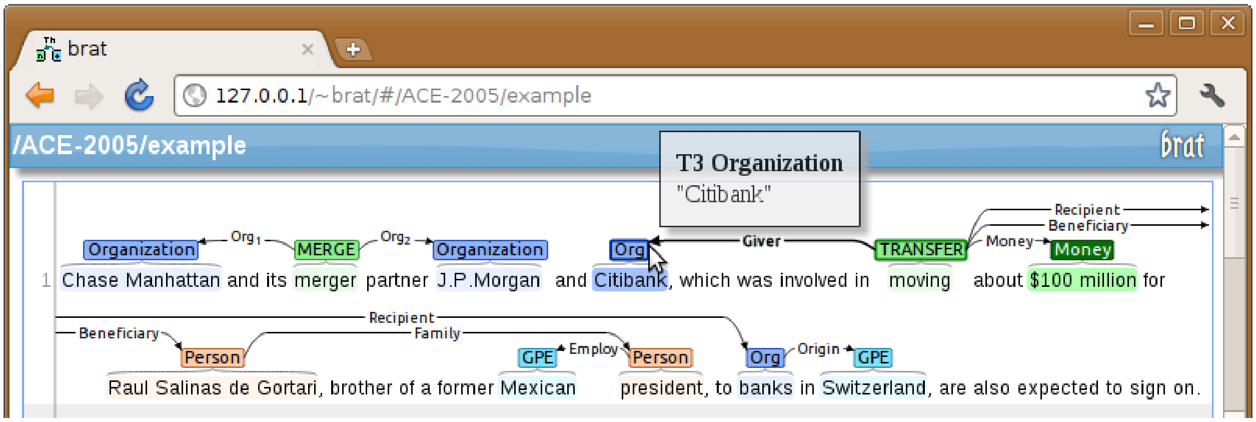
Figure 5. relation and event annotation guidelines

BRAT Annotation Capabilities

Various annotation tasks can be performed using BRAT such as

- Entity mention detection
- Event extraction
- Conference resolution
- Normalization
- Chunking
- Dependency syntax
- Meta-knowledge

BRAT can be used for many other tasks other than annotation such as **visualization** and **information extraction system evaluation.**

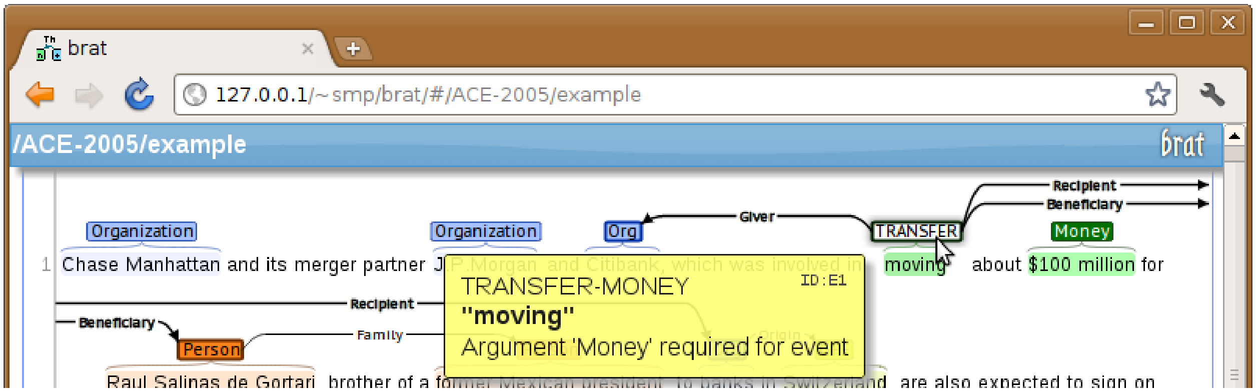


1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million for

Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

Figure 6. annotation visualization

BRAT supports Real-time collaboration using a client-server architecture which allows multiple annotators to work simultaneously on the corpora and be able to view each other's edits in real-time. It is also fully configurable allowing each document to have its own configuration, allowing multiple projects to be hosted on a single server. Annotation validation is included in BRAT, and it can check all the restrictions that can be provided in its expressive configuration.



1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million for

Raul Salinas de Gortari brother of a former Mexican president to banks in Switzerland are also expected to sign on

TRANSFER-MONEY
"moving"
Argument 'Money' required for event

Figure 7. incomplete annotation and details

Implementation

BRAT is built on a client-server architecture with HTTP and JSON as the communication protocol. It can easily be modified to replace the client or server. XHTML and SVG - Scalable Vector Graphics - are used by the client to connect with the server, which is implemented using JavaScript and XML (AJAX) technologies. It supports both CGI and FastCGI protocols and employs a stateless server back-end implementation in Python.

Other Tools

RADISP

RADISP - robust, accurate, but domain-independent, statistical parsing is a model which attempts to combine both phase boundaries and grammatical relations in the statistical annotation of general text. A beam search is done for the most probable overall analysis on the threshold output of each phase. In this model, the text is first tokenized and PoS and punctuation tagged using a HMM handling model. Next, deterministic morphological analysis and lemmatization is done on the tokens. Finally, the n-best parades are selected and displayed as syntactic trees and factored into a sequence of elementary predictions for a minimal recursion semantic representation.

DOCCANO

Doccano is an open-source annotation tool that provides text categorization, sequence tagging, and sequence-to-sequence operations. It can be used to build a dataset in a matter of hours. It used Python for the backend and JavaScript web app using Vue.js and Nuxt.js for its frontend.

Conclusion

We have stated the importance and use of NLP in text annotation, the various techniques used to develop annotation tools, and the features of a good annotation tool. We also reviewed BRAT which aims to increase annotator productivity using NLP technology along with a few other tools being developed and researched for annotation tasks.

BRAT with its conversion tools and documentation is freely available under MIT license from its homepage at http://brat.nlplab.org.

Acknowledgments

The authors of this article would like to thank the various authors of the research papers and other articles that were referred to in this review article. This work was supported by N Nasurudeen Ahamed, instructor In-Charge for Information Retrieval and Organization, Presidency University Bengaluru.

REFERENCES

[1] @inproceedings{Stenetorp2012bratAW,
title={brat: a Web-based Tool for NLP-Assisted Text Annotation},
author={Pontus Stenetorp and Sampo Pyysalo and Goran Topic and Tomoko Ohta and Sophia Ananiadou and Junichi Tsujii},
booktitle={EACL},
year={2012}
}

[2] S. bai, J. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling”, arXiv, April, 2018.

[3] P. Rémy, “Keras TCN”, GitHub https://github.com/philipperemy/keras-tcn, January. 2021.

[4] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, and L. E. Barnes, “Random Multimodel Deep Learning for Classification”, arXiv, April, 2018.

[5] https://dida.do/blog/the-best-free-labeling-tools-for-text-annotation-in-nlp

[6] @inproceedings{briscoe2002robust,
title={Robust accurate statistical annotation of general text.},
author={Briscoe, Ted and Carroll, John A},
booktitle={LREC},
year={2002}

}

[7] https://brat.nlplab.org/manual.html

[8] https://users.ox.ac.uk/~martinw/dlc/chapter2.htm

[9] A. Ramisa, F. Yan, F. Moreno-Noguer and K. Mikolajczyk, "BreakingNews: Article Annotation by Image and Text Processing," in IEEE Transactions on Pattern Analysis
and Machine Intelligence, vol. 40, no. 5, pp. 1072-1085, 1 May 2018, doi: 10.1109/TPAMI.2017.2721945.

[11] What is Language Data Annotation and how it is useful in Machine Learning & AI? | by Rayan Potter | Nerd For Tech | Medium

[12] 1. The Basics - Natural Language Annotation for Machine Learning [Book] (oreilly.com)

[13] https://github.com/doccano