Bureau for Rapid Annotation Tool: Collaboration can do More over Variety-oriented Annotations

Zheng Wang
wangz@istic.ac.cn
Institute of Scientific and Technical Information of China
Haidian District, Beijing, P. R. China

Shuo Xu*
xushuo@bjut.edu.cn
Beijing University of Technology
Chaoyang District, Beijing, P. R. China

Abstract

A high-quality manually annotated corpus is crucial for many text mining and information extraction tasks. Several workbenches have been developed in the literature to facilitate collaborative annotation. However, given the growing volumes of un-annotated documents, these variety-oriented annotation workbenches have many shortcoming in terms of teamwork, quality control and time effort. For this purpose, we develop a novel workbench such that collaboration can do more over variety-oriented annotation. Our workbench is named as Bureau for Rapid Annotation Tool (Brat for short). Main functionalities include enhanced semantic constraint system, Vim-like shortcut keys, annotation filter and graph-visualizing annotation browser. Until now, over 500,000 mentions have been annotated with our Brat workbench.

ACM Reference Format:
Zheng Wang and Shuo Xu. 2021. Bureau for Rapid Annotation Tool: Collaboration can do More over Variety-oriented Annotations. In *EEKE 2021 - Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

A high-quality manually annotated corpus is very crucial for many text mining and information extraction tasks [1, 3, 4, 6, 7, 14, 15]. Several workbenches have been developed in the literature to facilitate collaborative annotation [8, 11, 13, 16]. However, given the growing volumes of un-annotated documents, these variety-oriented workbenches still have many shortcomings in terms of teamwork, quality control and time effort. Let's take the sentence ”Depending on the model, a Tesla costs somewhere between 1 and 3.33 BTC” [2] as an example. A practical issue we face is whether or not to assign ”Price” type to the mention ”between 1 and 3.33 BTC”. This actually depends on a consensus acknowledging *Bitcoin* as actual money [5, 17].

Table 1. Two types of annotation collaborations used in previous workbenches.

Grounded	Trusted
UIMA-type system[10] Annotation semantic constraint [13]	Teamwork [8, 16] Personal workspace TeamTat [16] Multi-annotator analysis [8] Pairwise annotators comparison [8]

Reaching this consensus is extremely time-consuming and heavily rely on two types of annotation collaborations in Table 1: *grounded collaboration* and *trusted collaboration*. By grounded collaboration, we mean that the resulting annotators are restricted with sounded pre-arrangements. For example, U-Compare only supports named entity annotations in the UIMA-type system [10], which can avoid many conflicts. An alternative [13] takes the form of semantic constraints. In more details, a certain relationship should take parameters with specific entity types. As for trusted collaboration, YEDDA [16] recognized common gestures from BRAT [13] and embedded many functionalities including teamwork, multi-annotator analysis and pairwise annotators comparison. Then, on the basis of various annotations, the inter-project agreement can be calculated. Another strategy of trusted collaboration, user-independent workspace, was utilized in TeamTat [8].

This paper combines these two types of annotation collaborations to structure various mentions annotated by each annotator and develop a workbench named as Bureau for Rapid Annotation Tool (Brat for short). Main functionalities include enhanced semantic constraint system, Vim-like shortcut keys, annotation filter and graph-visualizing annotation browser. Until now, over 500,000 mentions have been annotated with our Brat workbench.

2 Functionalities

2.1 Enhanced Semantic Constraint System

It is well known that not all parameters are valid to a specific relationship. To limit invalid annotated results for an annotation project, its manager can customize the schema at

*Corresponding author

EEKE 2021, September 27–30, 2021, Online
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

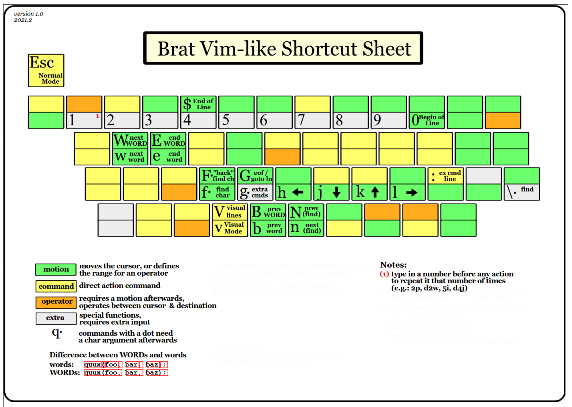

version 1.0
2021.2
Brat Vim-like Shortcut Sheet
Esc Normal Mode
1 (1)
2
3
$ End of Line
4
5
6
7
8
9
0 Begin of Line
W next WORD
w next word
E end WORD
e end word
F· "back" find ch
f· find char
G eof / goto ln
g· extra cmds
h ←
j ↓
k ↑
l →
: ex cmd line
\· find
V visual lines
v Visual Mode
B prev WORD
b prev word
N prev (find)
n next (find)
motion moves the cursor, or defines the range for an operator
command direct action command
operator requires a motion afterwards, operates between cursor & destination
extra special functions, requires extra input
q· commands with a dot need a char argument afterwards
Difference between WORDs and words
words: quux(foo, bar, baz);
WORDs: quux(foo, bar, baz);
Notes:
(1) type in a number before any action to repeat it that number of times (e.g.: 2p, d2w, 5i, d4j)

Figure 1. Vim-like Shortcut Keys Mapping

any time. Once the schema is modified, all involved annotated mentions will be adjusted correspondingly. A readable name is usually assigned to each type of entity and relation. In addition, a list of rules are also attached to expression the constraint conditions between parameters in each type of relation. In this way, the understanding on entities and relations from the manager can be delivered to all annotators.

2.2 Vim-like Shortcut Key

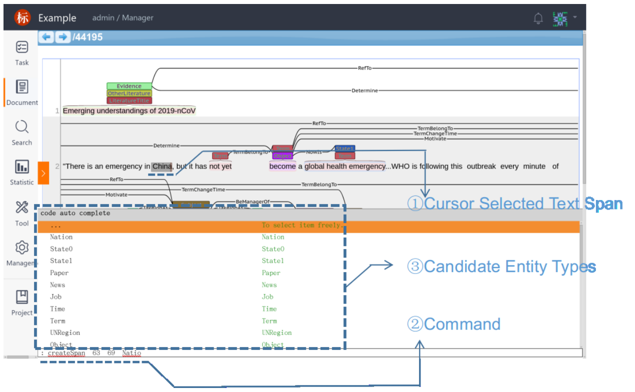
According to our observation, the conventional annotating operations (marking, selecting and confirming [13]) is time-expensive to choose a proper candidate from more than 5 entity types or relationships. To speed up the annotation procedure, our workbench embeds many Vim-like shortcut keys [12]. In this time, one can annotate smoothly an entity by the following steps (cf. Figure 2): 1) to move cursor and select a span of text with Figure 1, 2) to acknowledge one command from recommended candidates with TAB and ENTER, 3) to type leading characters and confirm entity types. Similar operations can be followed for relation mention annotation. It is worth noting that the key feature of this functionality is code auto-completion. This is based on enhanced semantic constraint system and polymorphic type inference [9].

2.3 Configurable Annotation Filter

It's unknown in advance how many mentions should be annotated for a single document, especially a very long document. To correct wrong annotations in time and reduce the conflicts among multiple documents, a feasible solution is to only display the mentions with interested types in current workspace. Thereupon, we provide a configurable annotation filter by toggling or un-toggling entity types and relationships.

2.4 Graph-visualizing Browser

In real-world scenario, it is not trivial to reach an agreement when multiple annotators are involved, and an entity or relation is mentioned simultaneously in multiple documents.



1 Emerging understandings of 2019-nCoV

2 "There is an emergency in China, but it has not yet become a global health emergency...WHO is following this outbreak every minute of

```
code auto complete
...                                   To select item freely.
Nation                                Nation
State0                                State0
State1                                State1
Paper                                 Paper
News                                  News
Job                                   Job
Time                                  Time
Term                                  Term
UNRegion                              UNRegion
Object                                Object
: createSpan 63 69 Natio
```

①Cursor Selected Text Span

③Candidate Entity Types

②Command

Figure 2. The novel annotation procedure powered by Vim-like Shortcut Keys

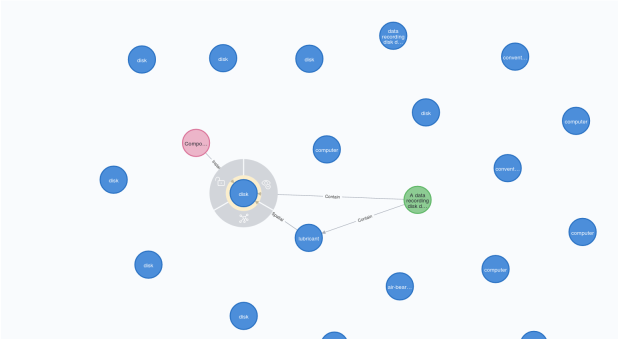


Figure 3. The visualization of the entity "Disk" and its relevant information in the TFH-2020 corpus [4] via our browser

To inspect the underling disagreements, our workbench can load and index all texts, mentions and their types, and then visualize them in a graph browser, as illustrated in Figure 3.

3 Conclusion

Many projects utilized our Brat workbench to annotate interested entities and/or relations, and inspect potential conflicts over variety-oriented annotations. Nowadays, over 500,000 mentions have been annotated with our Brat workbench. In the near future, the Vim-like shortcut keys will be strengthen further, and machine learning methods will be incorporated to accelerate conflict inspection.

Acknowledgments

This work is supported partially by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA16040504), National Key Research & Development Program of China (Grant No. 2019YFA0707202), and National Natural Science Foundation of China(Grant No. 71704169 and 72074014). We also thank Professor Yiming Jing and Rui Zheng for their assistance on how to understand the collaboration in the field of psychology.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. *Lecture Notes in Computer Science* 4825 LNCS (2007), 722–735.

[2] Jeff Benson. 2021. Here's How Much a Fully Loaded Tesla Model S Will Cost You in Bitcoin. https://decrypt.co/57071/heres-how-much-a-fully-loaded-tesla-model-s-will-cost-you-in-bitcoin [Online; accessed 16-Mars-2021].

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. 1247–1250.

[4] Liang Chen, Shuo Xu, Lijun Zhu, Jing Zhang, Xiao-ping Lei, and Guancan Yang. 2020. A Deep Learning based Method for Extracting Semantic Information from Patent Documents. *Scientometrics* 125, 1 (2020), 289–312.

[5] Vanessa Dirwai. 2021. Should Christians Trade Bitcoin And Other Cryptocurrencies? https://preciousearnings.medium.com/should-christians-trade-bitcoin-and-other-cryptocurrencies-878441e702c5 [Online; accessed 22-August-2021].

[6] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: a Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference*. 601–610.

[7] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open Information Extraction from the Web. *Commun. ACM* 51, 12 (2008), 68.

[8] Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. TeamTat: A Collaborative Text Annotation Tool. *CoRR* abs/2004.11894 (2020).

[9] Steven L. Jenkins and Gary T. Leavens. 1995. Polymorphic Type Inference in Scheme. *Computer Science Technical Reports* 75 (1995).

[10] Yoshinobu Kano, William A. Baumgartner Jr., Luke McCrohon, Sophia Ananiadou, K. Bretonnel Cohen, Lawrence Hunter, and Jun'ichi Tsujii. 2009. U-Compare: Share and Compare Text Mining Tools with UIMA. *Bioinform.* 25, 15 (2009), 1997–1998.

[11] Mariana L. Neves and Ulf Leser. 2014. A Survey on Annotation Tools for the Biomedical Literature. *Briefings Bioinform* 15, 2 (2014), 327–340.

[12] Kim Schulz. 2007. *Hacking Vim: A Cookbook to Get the Most Out of The Latest Vim Editor.* Packt Publishing Ltd.

[13] Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-Assisted Text Annotation. In *Conference of the 13th European Chapter of the Association for Computational Linguistics*, Walter Daelemans, Mirella Lapata, and Lluís Màrquez (Eds.). 102–107.

[14] Zheng Wang, Shuo Xu, and Lijun Zhu. 2018. Semantic Relation Extraction Aware of N-Gram Features from Unstructured Biomedical Text. *Journal of Biomedical Informatics* 86 (2018), 59–70. https://doi.org/10.1016/j.jbi.2018.08.011

[15] Shuo Xu, Xin An, Lijun Zhu, Yunliang Zhang, and Haodong Zhang. 2015. A CRF-based System for Recognizing Chemical Entity Mentions (CEMs) in Biomedical Literature. *Journal of Cheminformatics* 7, Suppl 1 (2015), S11. https://doi.org/10.1186/1758-2946-7-S1-S11

[16] Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *Proceedings of the 56th Annual Meeting Association for Computational Linguistics*, Fei Liu and Thamar Solorio (Eds.). 31–36.

[17] David Yermack. 2015. Chapter 2 - Is Bitcoin a Real Currency? An Economic Appraisal. In *Handbook of Digital Currency*, David Lee Kuo Chuen (Ed.). Academic Press, San Diego, 31–43.