Text Annotation Graphs (TAGs)

| Aditi Mallavarapu  University of Illinois at Chicago  Chicago, USA  amalla5@uic.ed | Aryadip Sarkar  University of Illinois at Chicago  Chicago, USA  asarka7@uic.edu | Hai Tran  University of Illinois at Chicago Chicago, USA  htran41@uic.edu |
| --- | --- | --- |

# ABSTRACT

Language hides some inherent relations among its constituents. Illustrating their relations to understand the relations though may seem trivial but has been considered to be very challenging. As NLP problems are being more complex and the amount of data is increasing rapidly, NLP softwares are expected to be robust and comprehensive. Aiding the analysts with proper tools to visualize and to analyze these complex relations is necessary. Having a good analysis tool would also help NLP expand to other technical domains such as biology, chemistry and many others. The tool that we have developed helps in analyzing the relationship among words and among words and other components of a sentence like phrases, or sentences etc. in the text by annotating the relations between these components.

## Author Keywords

Text annotations; graph visualizations; Natural language processing.

# INTRODUCTION

Annotation can often be time consuming and very domain specific. Although there are tools that would reduce the annotation time and cost, but certain relations get masked. Our effort was to design a tool that not only makes the annotation faster but makes the analysation after annotation more intuitive. We intend to elicit all the relationships in the text, not only the primary ones and be able to differentiate among them as per the domain specific needs. We designed two views to make the tool easier to use. The text- view which would have the text and the relationships with arrows spanning across the words. The second view is the staircase view which would represent each sentence as a series of word staircase and relationships with arrows from one stair to the other. We have used different texture and color coding to differentiate between the types of relationships and the relationships which span outside the sentence, to make it more intuitive. These two views can be used in conjunction to view and analyze the relationships for the text.

**related works**

Two existing tools that we know of are BRAT (+ its predecessor STAV) and Odin Open Domain Rule Visualizer. Both tools are fairly new for public use (BRAT was published in 2012 and Odin was 2015), therefore they are not yet optimized for NLP researchers. Although BRAT comes with a plenty of functionalities, it only supports node-to-node (indicating connection between words or phrases to other words or phrases respectively) connections. We plan to support node-to-edge words to phrases), edge-to-node (phrases to words), and edge-to-edge ( phrases to phrases) connections as well. Another problem with BRAT is to identify the end connections. It is cumbersome to see the other end of the connection when the connection spans across many lines.

Although ODIN was able to address this issue the user has to encode the relationships in a rule based language, which might not be very friendly to scientists who deal with only domain specific stuff. Drawing on this, there is a need for a tool that would be able to elicit the relationships and also make the tool user friendly for the domain experts and domain independent to be specific.

**Analysis Tasks**

We are using two JSON files:

edge.json which gives us the specifications about a certain edge.

Eg: "Id":"E1","sourceid":"T44","destinationid":"T9","label":"Theme"

ID represents the ID of a specific relationship. Let’s consider a relationship between two words. Sourceid is the source word from where the relationship starts and destinationid is the ending word where the relationship ends. Label is the name of the relationship through which both are connected. Hence in the above example, the source word is at T44 (44th word of the document) is connected to T9 (9th word of the document) through the relationship “Theme” and the ID of this relationship is given by E1. This is the case of word to word relationship.

"Id":"E9",   "sourceid":"T45",   "destinationid":"E8",   "label":"Cause"}]

Similarly in the above example, the source word is T45(45th word in the document) which is connected to the to E8(the 8th edge in the document) through the relationship tagged as Theme and it’s given an ID of E1. This is the case of word to edge relationship.

2) labelleddata.json which given the identification of every word in the document

Eg: "Id": "T1","wordindex": 1,"word": "acid"

The position of the word in the document is given by the ‘wordindex’ and the word is given the tag ‘word’ and the Id of the word is given by ‘Id’.In the above example, the position of the word ‘acid’ in the document is 1 and the id of the word is T1.

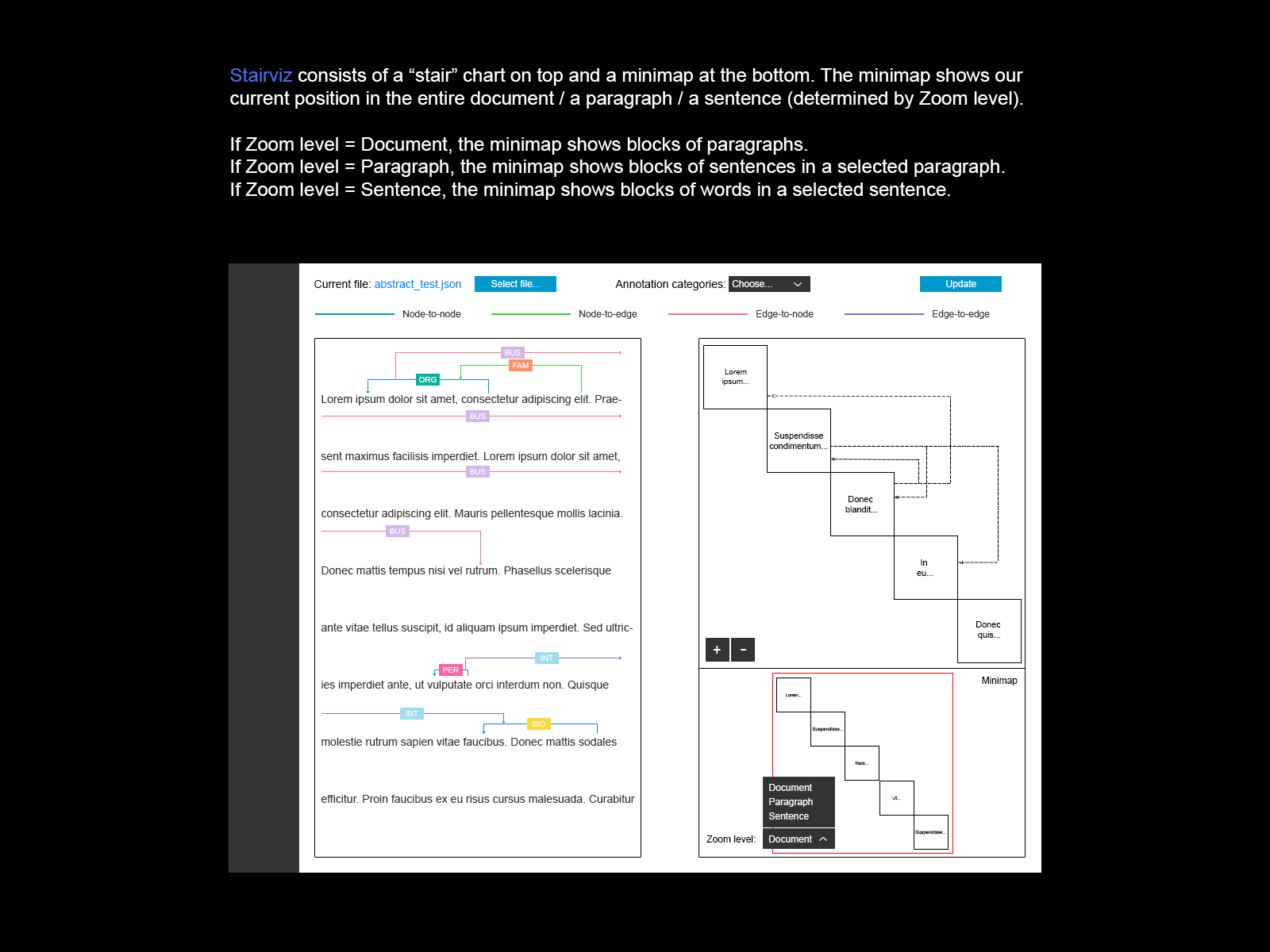


Figure 1. The mock design of the Text annotation Graph visualization containing the text view ( left) and the Staircase view (right)

**Case study**

**Expert feedback**

The experts were thoroughly impressed by the design especially the stair-case diagram. They really liked the idea which they said , like they haven’t seen that in context of text annotation.

They expect cross sentence and cross paragraph relationship in the future, so that will be pretty challenging with the recent architecture.

**Conclusions and future work**

We wish to automate this work by enabling the user to choose the text file and elaborate on the relationships in simple english language independent of the domain. The application would take these files and internally convert them into a format suitable for the tool to visualize any text. Further we would want to enable the user to annotate new relationships by simply using the either the staircase view or the text - view.

# ACKNOWLEDGMENTS

Sample text: We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234-2012-ABC). This is just an example.

# REFERENCES

1. @\_CHINOSAUR. 2014. VENUE IS TOO COLD. #BINGO #CHI2016. Tweet. (1 May, 2014). Retrieved February 2, 2014 from https://twitter.com/\_CHINOSAUR/status/461864317415989248
2. ACM. How to Classify Works Using ACM’s Computing Classification System. 2014. Retrieved August 22, 2014 from [http://www.acm.org/class/how\_to\_use.html](http://www.acm.org/class/how_to_use.html%20)
3. R. E. Anderson. 1992. Social impacts of computing: Codes of professional ethics. *Soc Sci Comput Rev* 10, 2: 453-469.
4. Anna Cavender, Shari Trewin, Vicki Hanson. 2014. Accessible Writing Guide. Retrieved August 22, 2014 from <http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/>
5. Morton L. Heilig. 1962. Sensorama Simulator, U.S. Patent 3,050,870, Filed January 10, 1961, issued August 28, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal Ubiquitous Comput*. 18, 4 (April 2014), 765-766. <http://dx.doi.org/10.1007/s00779-014-0773-4>
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where do web sites come from?: capturing and interacting with design history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 1-8. <http://doi.acm.org/10.1145/503376.503378>
8. Nintendo R&D1 and Intelligent Systems. 1994. *Super Metroid*. Game [SNES]. (18 April 1994). Nintendo, Kyoto, Japan. Played August 2011.
9. Psy. 2012. Gangnam Style. Video. (15 July 2012.). Retrieved August 22, 2014 from <https://www.youtube.com/watch?v=9bZkp7q19f0>
10. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press.
11. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.
12. Langdon Winner. 1999. Do artifacts have politics? In *The Social Shaping of Technology* (2nd. ed.), Donald MacKenzie and Judy Wajcman (eds.). Open University Press, Buckingham, UK, 28-40.

**Columns on the last page should be of approximately equal length. Remove this line before submission.**