November 4, 2019

Discourse-Sentiment Alignment Tool (DSAT)

CPSC 547 Project Proposal Report

Patrick Huber (huberpat@cs.ubc.ca)

Introduction

Discourse parsing is a crucial task within the area of Natural Language Processing (NLP), which has shown to enhance many downstream applications, such as sentiment analysis, summarization and question answering. While many tasks within and outside of NLP have been greatly improved using novel deep learning methodologies (such as neural networks), discourse parsing itself could not yet take full advantage of deep learning methodologies, due to the limited amount of available data.
To overcome this data sparsity problem, I have been researching on an approach to automatically derive discourse structures from an auxiliary sentiment dataset using distant supervision for the past year. With this novel and completely automated approach to create semantic discourse trees, proposed in a recent paper [1], the tedious and expensive manual annotation by human linguistic experts has become obsolete. However, using this new methodology to create discourse structures from sentiment data, the human component is completely taken out of the loop. This calls for an implementation of an information visualization system, which allows users to explore the automatically generated discourse structure dataset, to be able to draw conclusions regarding the quality and alignment of the discourse structures with the given gold-standard sentiment.

Data and Task

The proposed project is within the domain of Natural Language Processing (NLP). The task itself is concerned with discourse parsing, a fundamental task within the area, which is mostly working with either complete or shallow discourse trees, representing the structure of a document (Figure 1). The available data for the project is generated in a previous research project published at the conference for Empirical Methods in Natural Language Processing (EMNLP), containing over 10,000 distinct discourse trees, each representing a document ranging from 2 to 150 atomic elements, so called elementary discourse units (EDUs). While there has been previous work [2] to compare different discourse trees generated by modern discourse parsing systems and compared against a common, available gold-standard tree, the combination of discourse structure trees and sentiment (but no available gold-standard data) is novel and has therefore not been explored before.

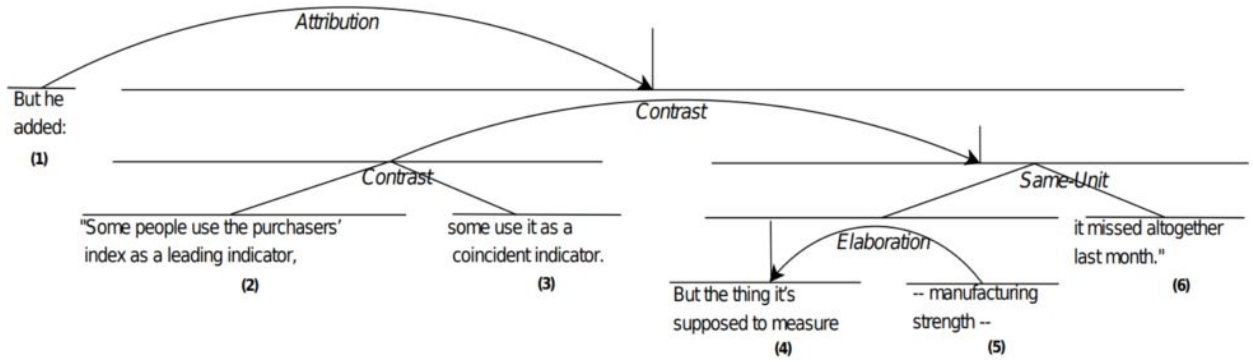

Figure 1: RST-style discourse tree

I am planning to use the available data to approach the task of creating an information visualization system to analyse and explore the automatically generated discourse trees regarding (1) a valid general tree structure (2) a well-aligned sentiment assignment for EDUs and resulting sub-trees (3) reasonable relative importance scores along the tree branches. The data contains a hierarchical textual structure representing the discourse tree, where every node in the tree (no matter if the node refers to a textual unit (EDU) or is an interior sub-tree) contains three values: (a) the sentiment score of the sub-tree, (b) the importance score of the sub-tree (c) the text covered by the node.
As the task is framed as an exploratory analysis of the data, the focus of the project lies in inspecting individual trees, rather than comparing them.

Proposed Infovis Solution

The proposed Infovis Solution to tackle the problem is shown in Figure 2.

The proposed interface thereby has three components:

- On the left: A document list with all possible discourse trees, which can be selected (additional highlightling showing the overall alignment of the tree sentiment with the gold-standard sentiment to find outliers easier)
- In the middle: The discourse tree/sentiment alignment represented as a node-link diagram, as the structure of the tree plays a crucial role in the visualization
- On the right: The document content numbered by EDU-number, which is aligned with the tree leave nodes (here: EDUs 1-17)

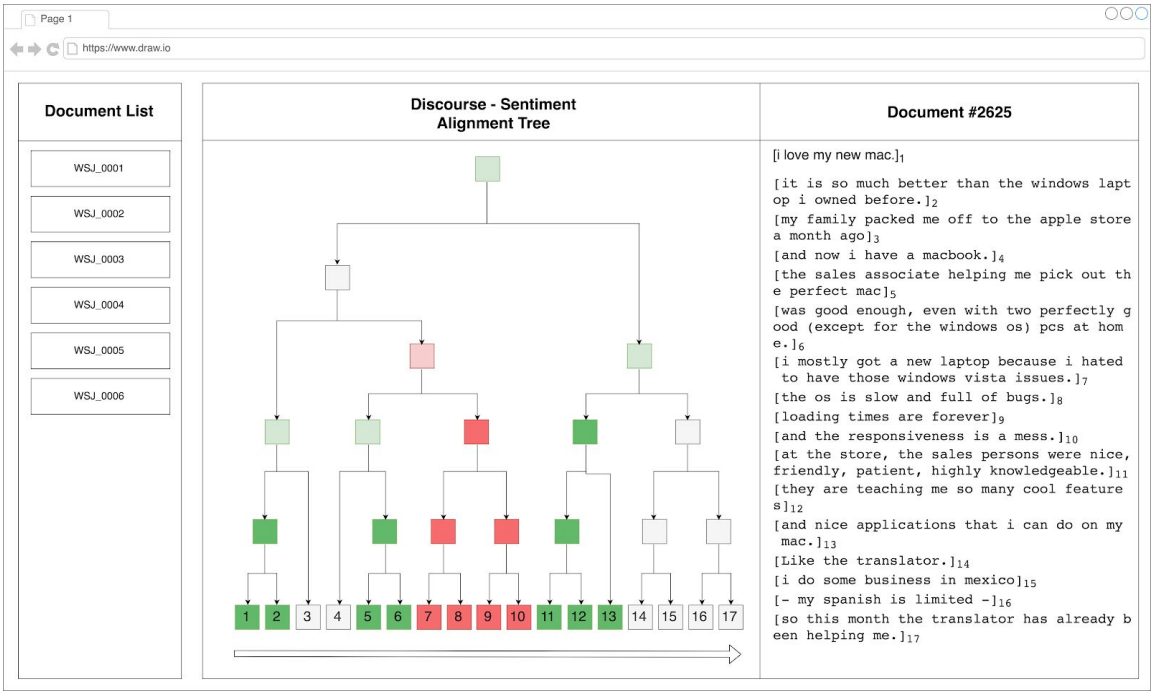

Figure 2: Overview of the Discourse-Sentiment Alignment Tool

Regarding the interaction idiom:
The discourse tree in the center and the discourse content on the right are proposed to be linked bidirectional. Furthermore, the center view should be zoomable using an idiom which ensures the relevant leaf-nodes to be always within the view, as only internal nodes do not carry any meaningful information without the related leaf-nodes.

Scenario of use:
In the standard usage scenario of the system the user selects a document in the selection on the left, which displays the relevant data of the document in the center and right panel. The user can then explore the sentiment augmented tree in the middle panel and dynamically select a sub-tree to focus her attention on the selected tree part. The right panel will dynamically adapt and highlight the selected subset of EDUs by reducing the opacity of not selected text fragments. Hovering over individual nodes or text fragments will highlight the matching selection in the other panel. This way, the user can find misaligned trees or wrong sentiment assignments and subsequently enhance the data or adapt the generation algorithm to account for those cases.

Implementation Approach

The proposed system will be implemented from scratch using the D3 framework [3]. The system will run in a browser window to make it easily accessible.

Milestones and Schedule

Task	Due Date	Est. Hours of Work	Description
Project Pitch		6h	Create slidedeck, practice presentation
Data Preprocessing	Oct 29	3h	The available dataset is in a task-specific textual format and needs to be transferred into a hierarchical data structure to be used for the visualization. The dataset transformation will be executed as a preprocessing step rather than an online computation, as the computational overhead would slow down the visualization
Creating the basic website layout	Oct 31	2h	Create the basic webpage setup, import necessary libraries and make the website responsive
Project Proposal	Nov 2	8h	Create mockup graphics, write report
Access the dataset through D3	Nov 1	2h	Load the external data through D3 and create some first outputs to get familiar with the structure and processing within D3
Create dataset explorer	Nov 8	2h	Bind the dataset to the document explorer on the right and display a button for each available document
Create tree visualization	Nov 15	10h	Create a dynamically scaled tree visualization with sentiment and importance scores encoded in tree nodes and links
Create the discourse panel and link panels	Nov 19	6h	Create the right view to explore the documents through the EDUs and link the selection of EDUs with the nodes in the tree
Allow selection of relevant sub-trees	Nov 22	6h	Allow the user to zoom into sub-trees by keeping relevant leaf-nodes in focus. Add “un-zoom” functionality

Add additional information channel to textual data	Nov 27	4h	Add glyph to discourse data to indicate EDU sentiment. Add additional indicator regarding overall sentiment alignment to document explorer
Final Presentation	Dec 10	15h	Create slidedeck, practice talk, demo rundown
Final Report	Dec 13	15h	Write report

Previous Work

[1] Huber, P. and Carenini, G., 2019. Predicting Discourse Structure using Distant Supervision from Sentiment. EMNLP 2019

[2] Zhao, J., Chevalier, F., Collins, C. and Balakrishnan, R., 2012. Facilitating discourse analysis with interactive visualization. IEEE Transactions on Visualization and Computer Graphics, 18(12), pp.2639-2648.

[3]Bostock M, Ogievetsky V, Heer J. D³ data-driven documents. IEEE transactions on visualization and computer graphics. 2011 Nov 3;17(12):2301-9.