Learning Noisy Discourse Trees

.

Boris Galitsky
Oracle Corp., Redwood Shores *CA USA*
boris.galitsky@oracle.com

Can discourse-level analysis help process noisy text?

- User-generated content is a noisy one, and for processing noisy data certain degree of abstraction and ascent to a higher-level of analysis seems to be beneficial
- Although discourse parsers rely on syntactic information, we expect them to perform reasonably well even when this information such as POS tags, NER and syntactic trees are incomplete and noisy
- To further overcome noisy text problem, we augment discourse trees with speech acts extracted from text to better represent the structure of what UGC authors communicate and in which way

If text is shorter than a paragraph, such as Twitter, discourse-level analysis is believed to be inappropriate

Current applications of Rhetoric Structure Theory

- There are some classes of NLP applications that are expected to leverage informational structure of text.
- DT can be very useful is text summarization.
- Knowledge of salience of text segments, based on nucleus-satellite relations proposed by (Sparck-Jones 1995) and the structure of relation between segments should be taken into account to form exact and coherent summaries.
- One can generate the most informative summary by combining the most important segments of elaboration relations starting at the root node.
- DTs have been used for multi-document summaries (Radev 2000).

Where we applied Discourse Trees

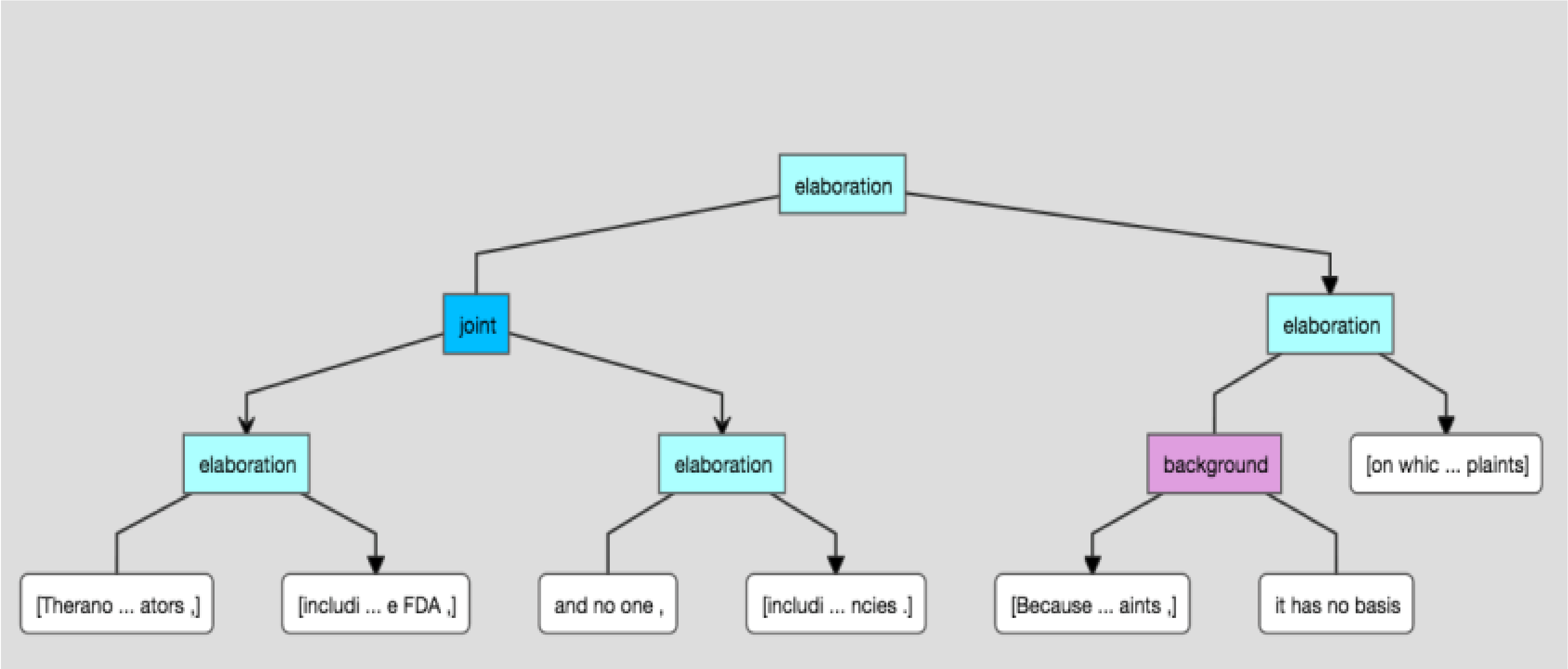
- Search engine: filtering out answers where keywords occur in wrong discourse units (ACL 2015)
- Content Generation (COLING 2016)
- Chat bots (EACL 2017)
- Finding answers with good rhetoric agreement and dialogue management (Oracle project)
- Finding an optimal sequence, given a set of paragraphs to form a cohesive text
- Discourse-level sentiment analysis
- Argumentation mining
- Document style recognition
- Author/source identification
- Text authenticity analysis

In all these areas, the case of noisy user-generated content is important and needs special treatment

What is a Discourse Trees

- Rhetoric Structure Theory (RST) simulates text organization by means of relations that hold between parts of text.
- RST explains coherence of text by forming a hierarchical, connected structure of texts, called Discourse Trees.
- Rhetoric relations are coordinate and subordinate ones that that hold across two or more text spans and therefore implement coherence.
- These text spans are called elementary discourse units (EDUs). The leaves of a Discourse Tree correspond to EDUs.
- Adjacent EDUs are connected by coherence relations (e.g., *Attribution, Sequence*), forming higher-level discourse units.

Theranos remains actively engaged with its regulators, including CMS and the FDA, and no one, including the Wall Street Journal, has provided Theranos a copy of the alleged complaints to those agencies. Because Theranos has not seen these alleged complaints, it has no basis on which to evaluate the purported complaints.



These application areas for noisy user-generated content

- *Sentiment analysis*. Sentiments can be inferred from such paragraph-level features as intense argumentation, complex mental states such as deception, and others.
- *Content validity* (authenticity, soundness, proper communication). We differentiate between valid, sound complaints requiring attention from invalid, fake ones where a user is in a bad mood or just intends to receive a compensation.
- *Commenting on a user post*. This is a special case of question answering, an automated support of user conversation, where the seed (the question or a request) is an incomplete or grammatically incorrect paragraph of text.

Customers reviews and opinionated text is a good source of data to explore how sentiment polarity can be inferred from the discourse-level features.

It is rather hard to assess the style features of a grammatically incorrect text based on its syntactic features. The degree of grammar deviation from normal is not a good indicator of content validity. .

To support a dialogue, a chat bot needs to extract a topic from a seed and also maintain the logical, rhetoric agreement between the seed and response.

Noisy text is classifier into two classes:

Positive (sentiment, correct / valid text, correct answer or reply);

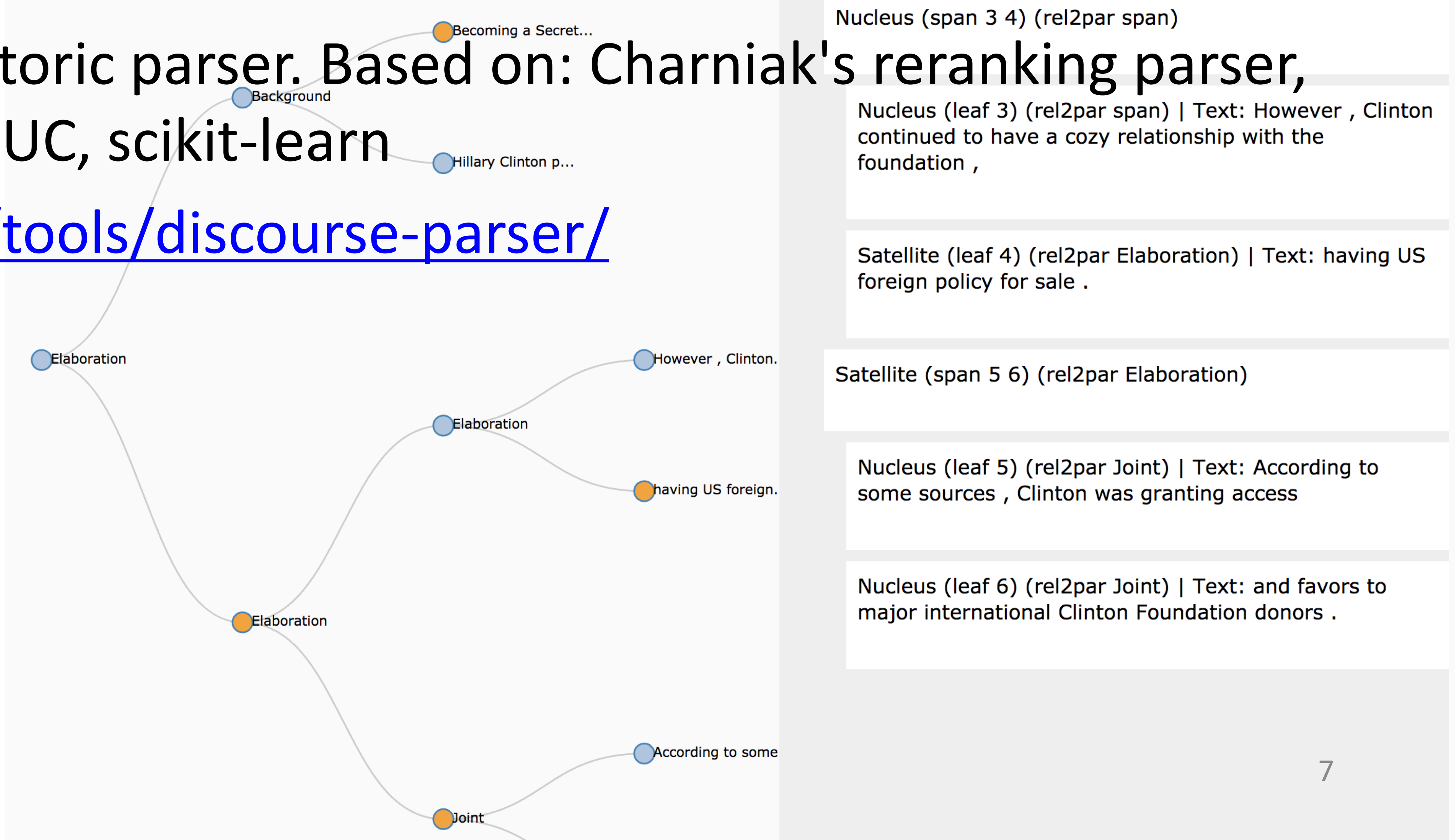
Negative (sentiment, incorrect / invalid text, incorrect answer or reply.

Sources of Discourse Trees

- (Surdeanu et al 2013) Rhetoric parser. Based on Stanford NLP http://agathon.sista.arizona.edu:8080/discp/comparison/fetch. It is integrated into our open source system https://github.com/bgalitsky/relevance-based-on-parse-trees

- (Joty 2015) Rhetoric parser. Based on: Charniak's reranking parser, Taggers from UIUC, scikit-learn http://alt.qcri.org/tools/discourse-parser/

Online tool:



Understanding discourse of a question

What does Clinton foundation do?

vs

What does Clinton foundation *really* do?

Most of the Clinton Foundation spending goes directly to programs that improve people's lives around the world

Vs

Becoming a Secretary of State, Hillary Clinton promised to distance herself from the Clinton Foundation. However, Clinton continued to have a cozy relationship with the foundation, having the US foreign policy for sale there. According to some sources, Clinton was granting access and favors to major Clinton Foundation donors

Question, clarification and answer

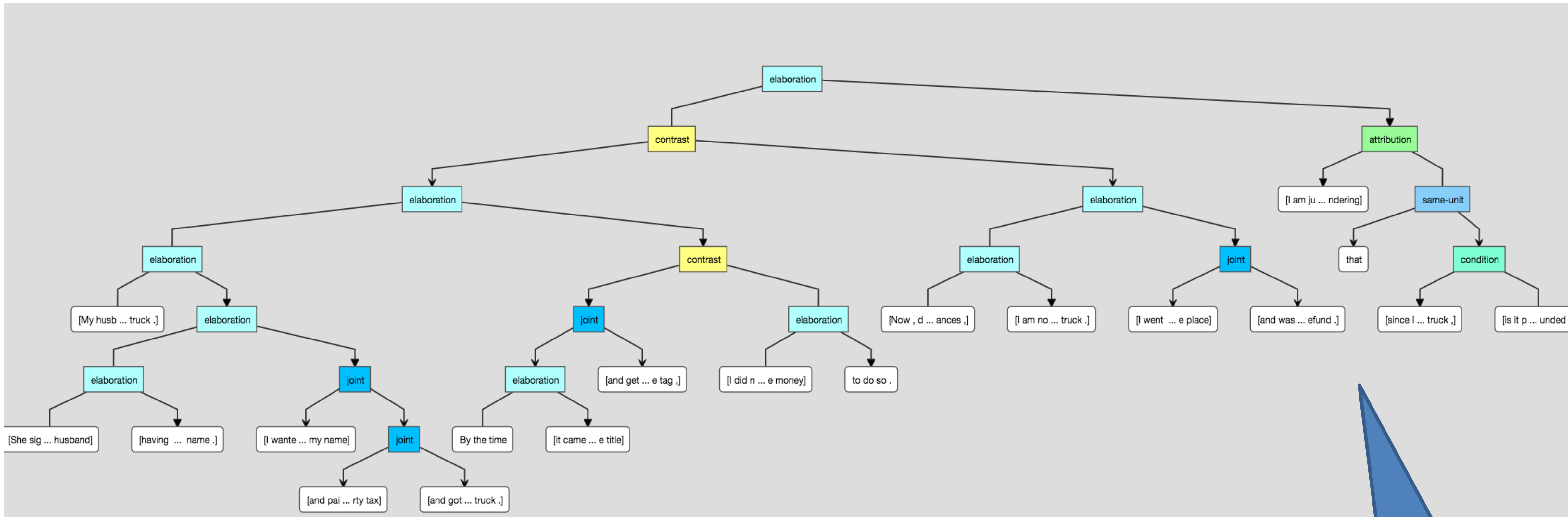
My husbands' grandmother gave him his grandfathers truck. She signed the title over but due to my husband having unpaid fines on his license, he was not able to get the truck put in his name. I wanted to put in my name and paid the property tax and got insurance for the truck. By the time it came to sending off the title and getting the tag, I didn't have the money to do so. Now, due to circumstances, I am not going to be able to afford the truck. I went to the insurance place and was refused a refund. Since I am not going to have a tag on this truck, can I get the property tax refunded?

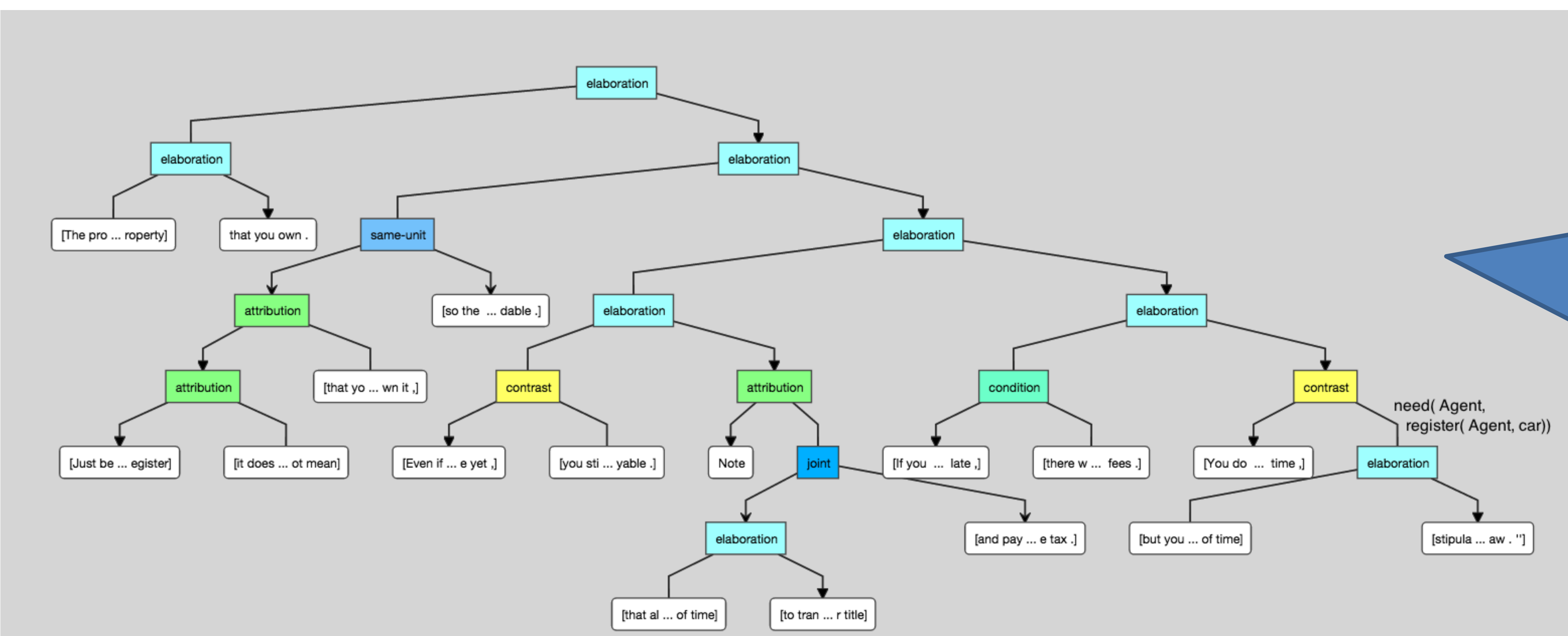
Are u talking about property tax refund, registration or insurance?

Insurance

The property tax is assessed on property that you own. Just because you chose to not register it does not mean that you don't own it, so the tax is not refundable. Even if you have not titled the vehicle yet, you still own it within the boundaries of the tax district, so the tax is payable. Note that all states give you a limited amount of time to transfer title and pay the use tax. If you apply late, there will be penalties on top of the normal taxes and fees. You don't need to register it at the same time, but you absolutely need to title it within the period of time stipulated in state law.

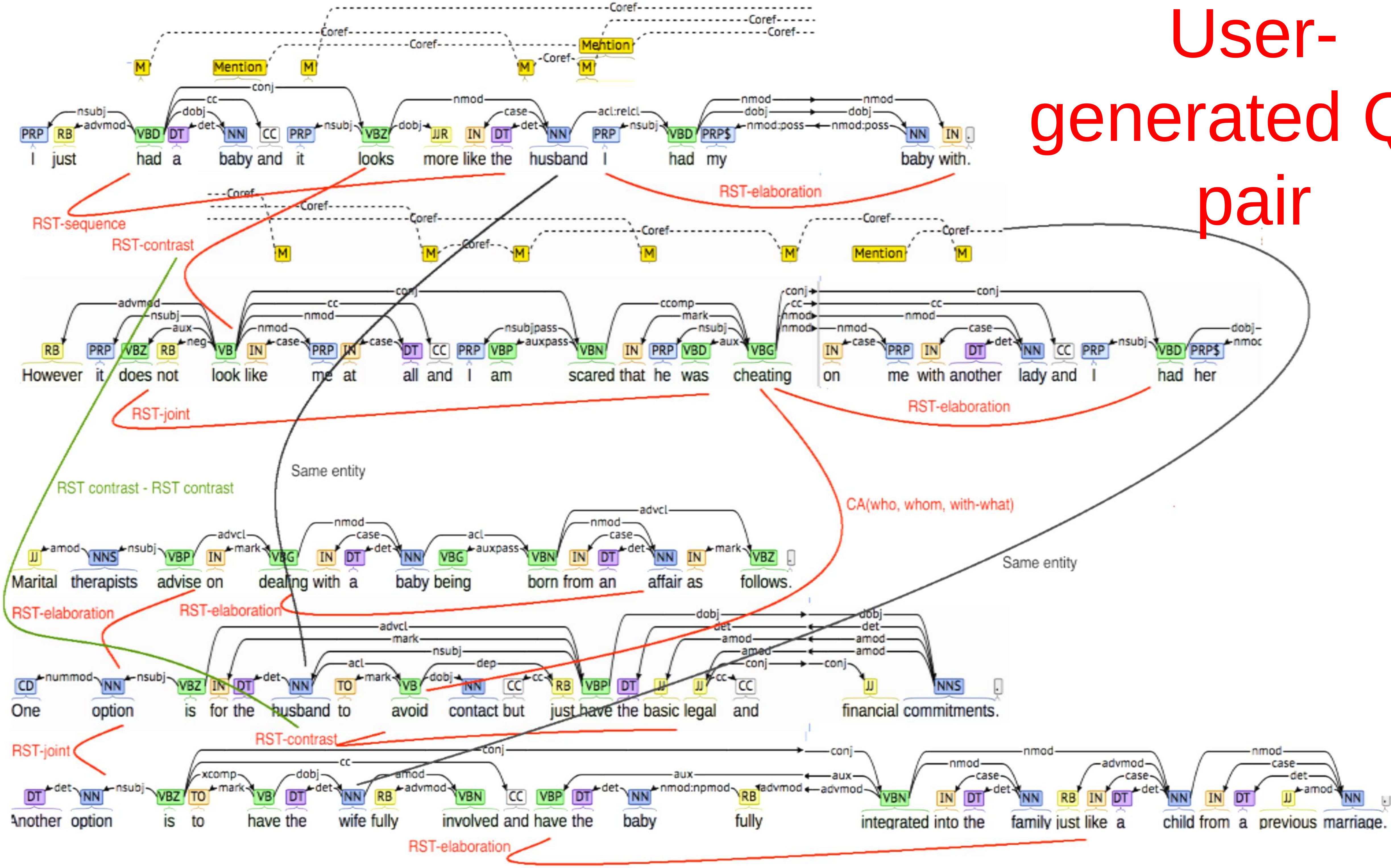
Coordination between a Q and A





Once relevance is established, coordination of Discourse Trees of Q and A is important to select an answer with suitable style

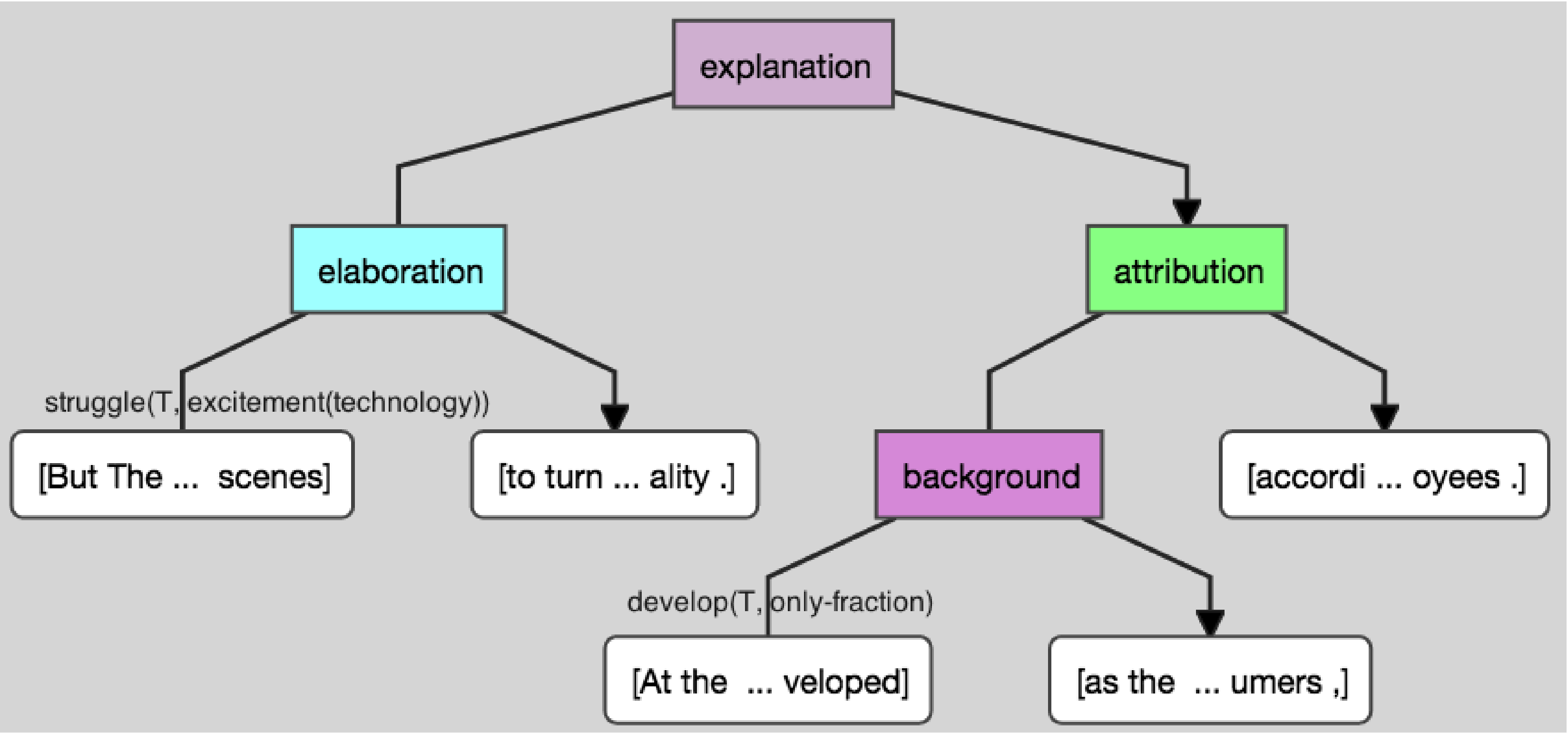
User-generated Q/A pair



Q: I just had a baby and it looks more like the husband I had my baby with. However it does not look like me at all and I am scared that he was cheating on me with another lady and I had her kid. This child is the best thing that has ever happened to me and I cannot imagine giving my baby to the real mom.

A: Marital therapists advise on dealing with a child being born from an affair as follows. One option is for the husband to avoid contact but just have the basic legal and financial commitments. Another option is to have the wife fully involved and have the baby fully integrated into the family just like a child from a previous marriage.

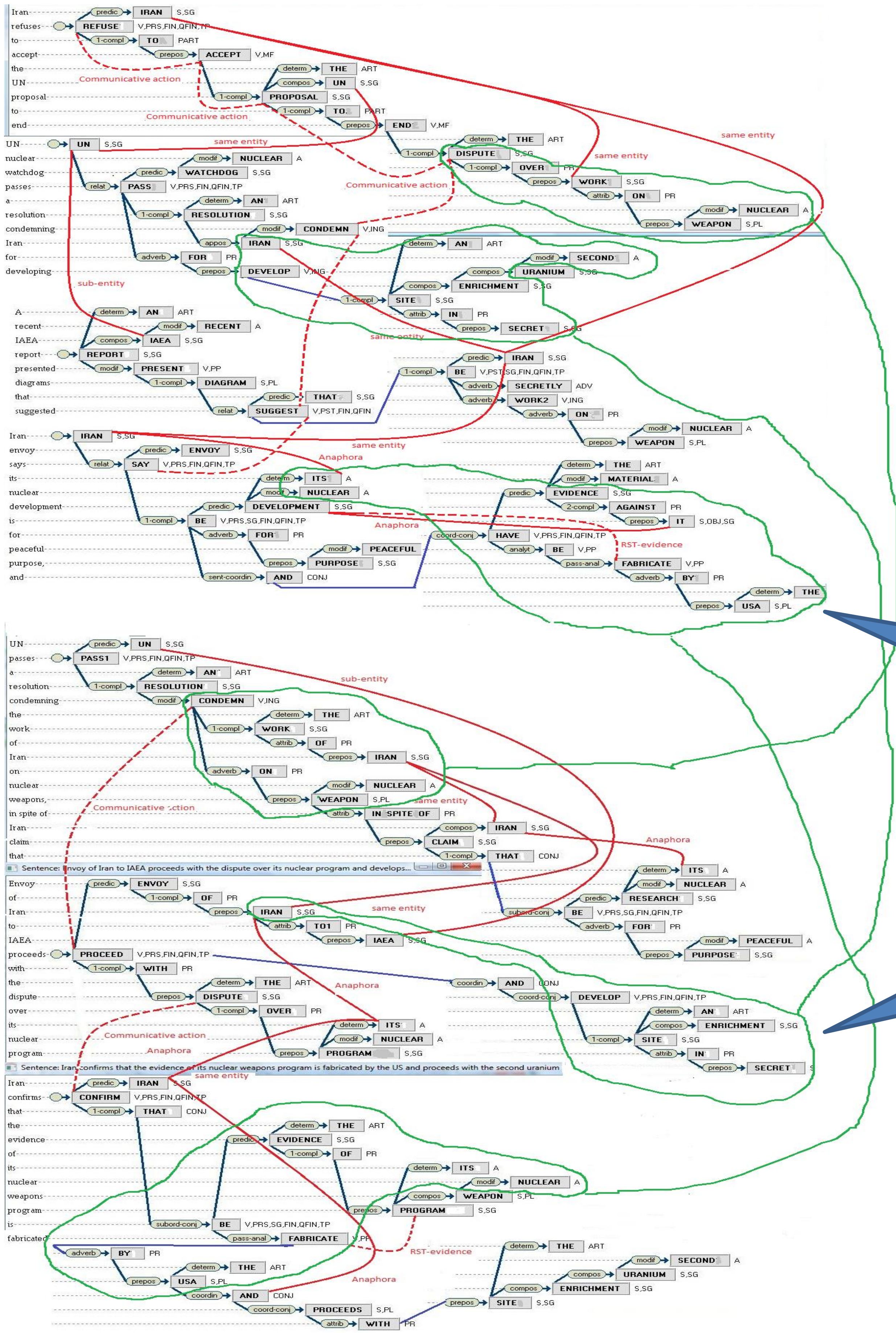
Extending Discourse Trees



"...But Theranos has struggled behind the scenes to turn the excitement over its technology into reality. At the end of 2014, the lab instrument developed as the linchpin of its strategy handled just a small fraction of the tests then sold to consumers, according to four former employees."

When arbitrary communicative actions are attached to a discourse tree as labels of its terminal arcs, it becomes clear that the author is trying to bring her point across and not merely sharing a fact.

Parse Thickets: complete representation for syntactic + discourse structure



Links for rhetoric relations might compensation for errors in syntactic parsing when computing similarity between two noisy texts

Scientific documents and their specific rhetoric structures

Let us compare the rhetoric structures of:

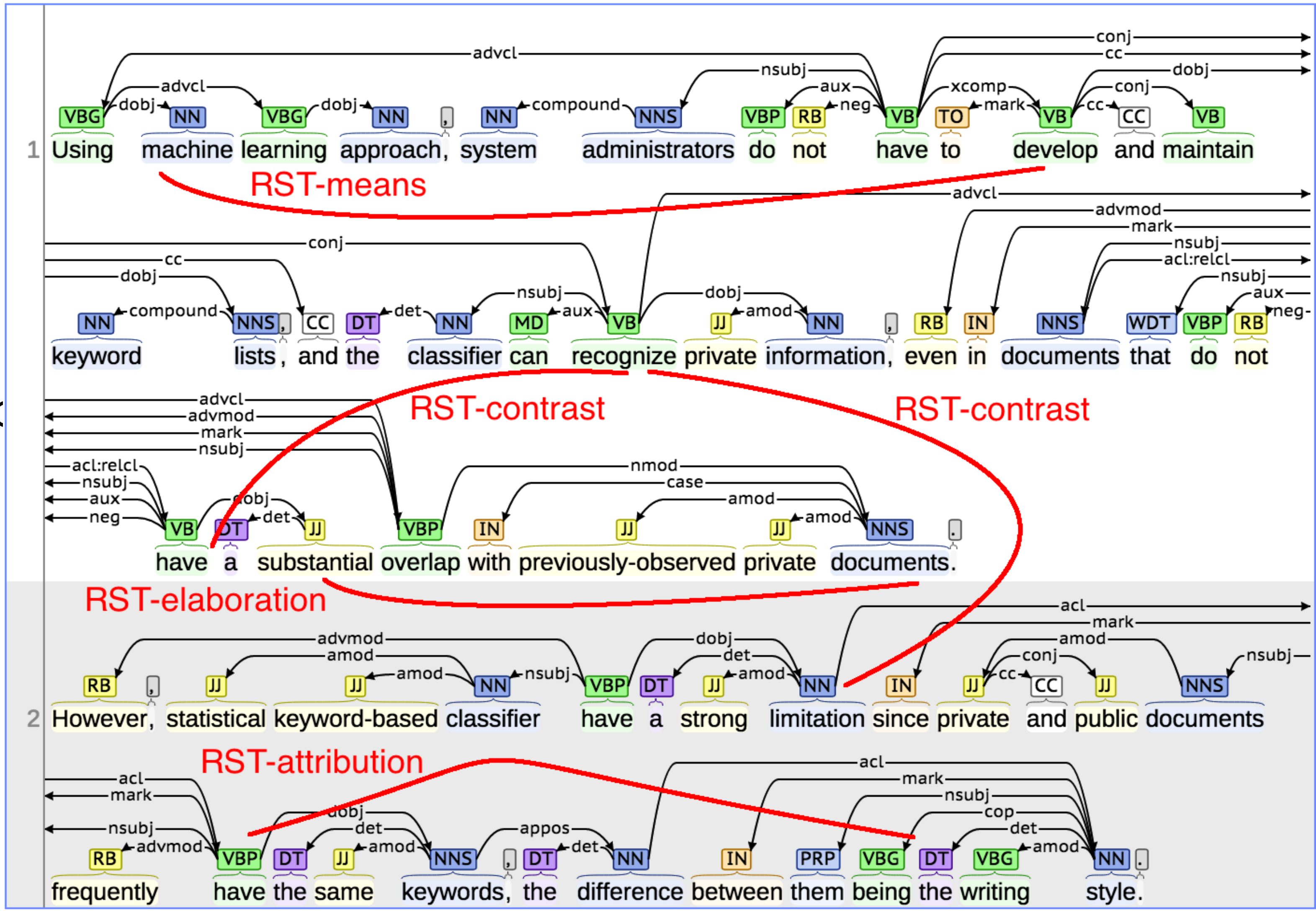
1) engineering system description text. Rhetoric relations include *sequence, purpose* and *elaboration*

2) scientific text outlining an state of knowledge in a given domain. Rhetoric relations include *means, contrast, attribution* and also *elaboration*

Describing a system, authors usually do not express their thoughts via contract or attribution, which is typical for a scientific discourse.

Other categories of sensitive documents:

- Legal: legally-binding contracts and agreements, offer letters, stock option certificates, letters of intent;
- Financial: plans, forecasts, internal rep
- Health: patient forms, requests, clinical data.
- Educational records: grades.

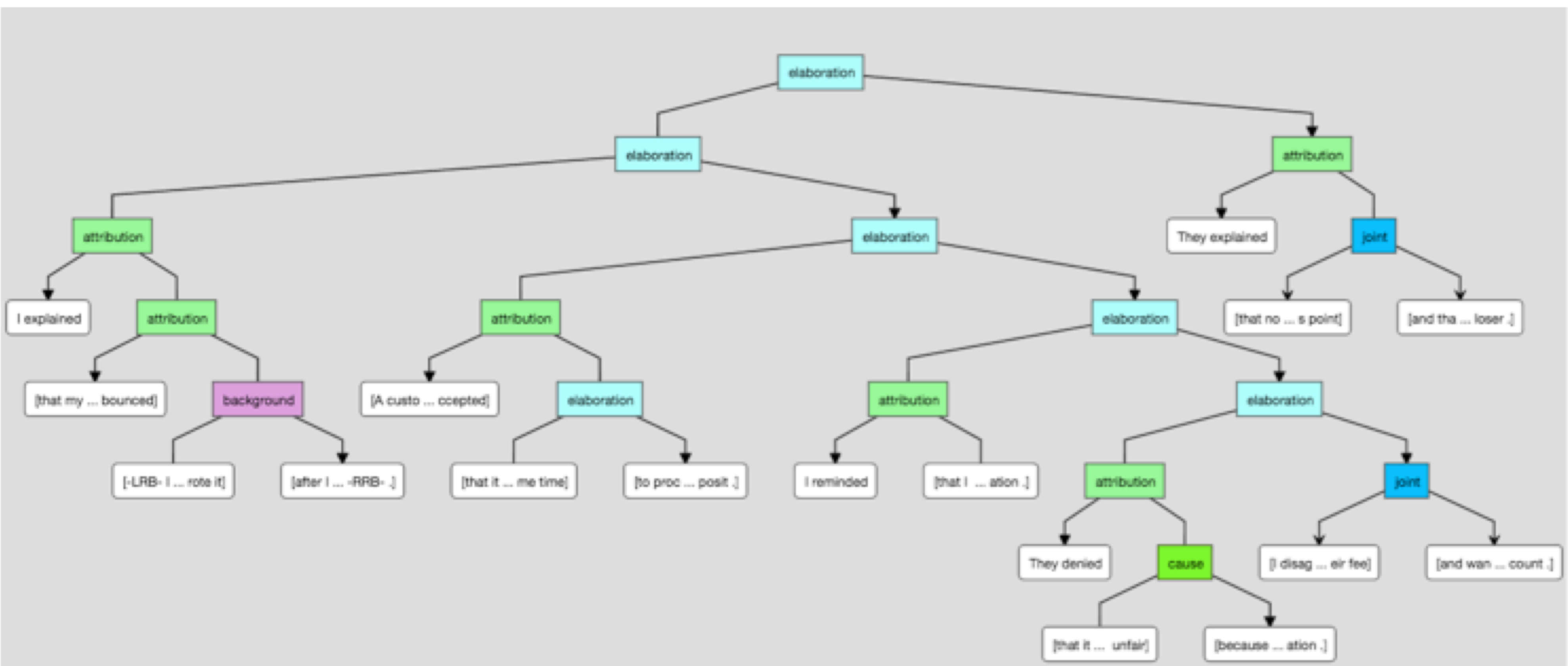


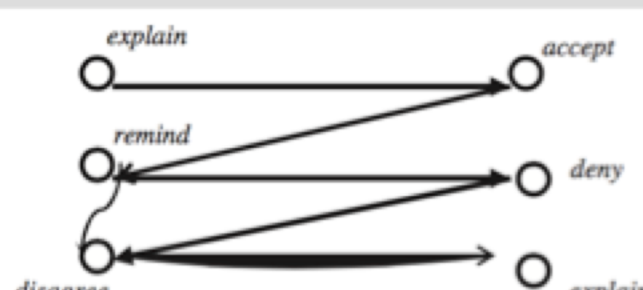
Discourse Representation of Interaction between Agents

*I **explained** that my check bounced (I wrote it after I made a deposit). A customer service representative **accepted** that it usually takes some time to process the deposit.*

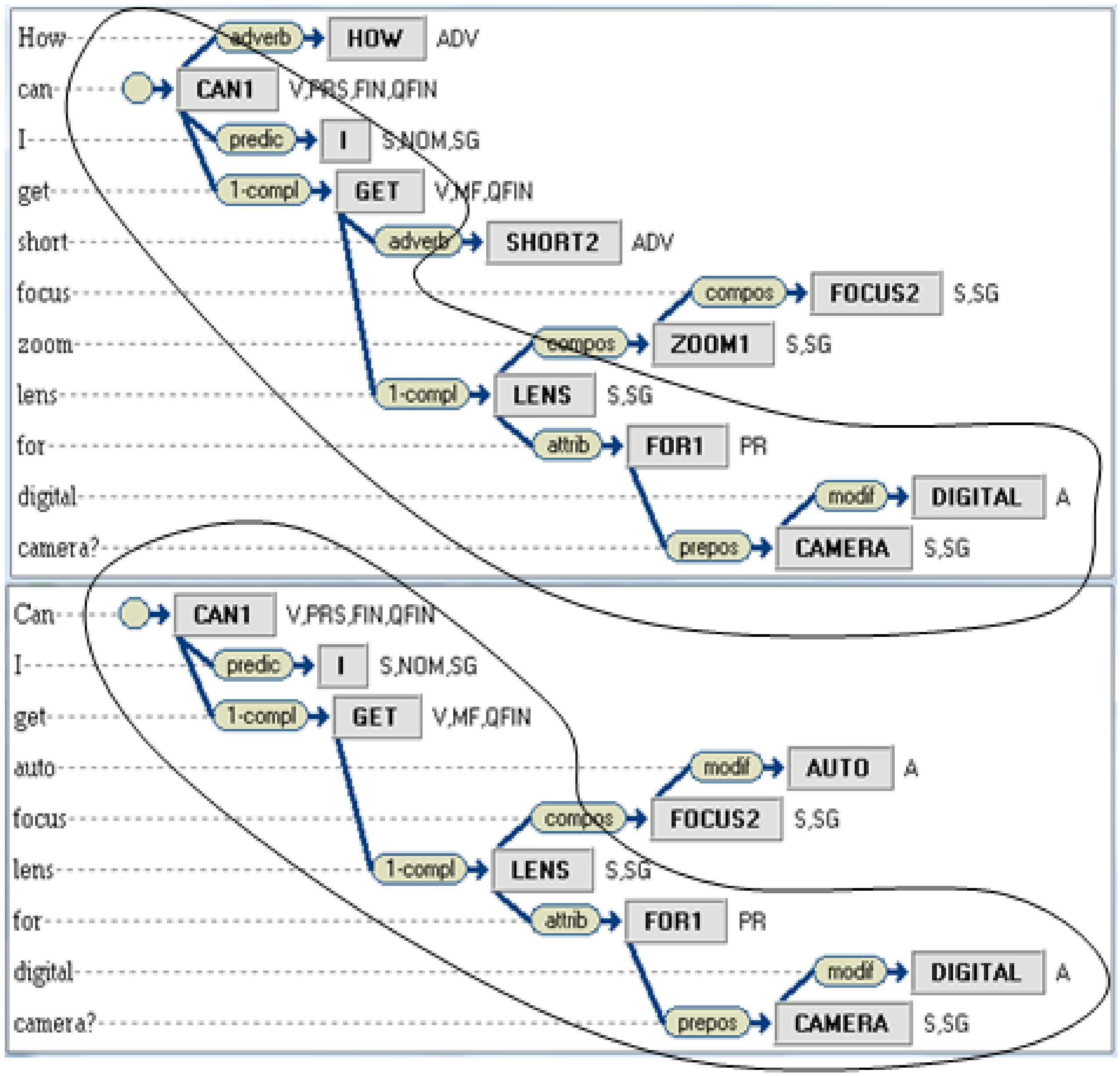
*I **reminded** that I was unfairly charged an overdraft fee a month ago in a similar situation.They **denied** that it was unfair because the overdraft fee was disclosed in my account information.*

*I **disagreed** with their fee and wanted this fee deposited back to my account. They **explained** that nothing can be done at this point and that I need to look into the account rules closer.*





Similarity = common part among parse trees



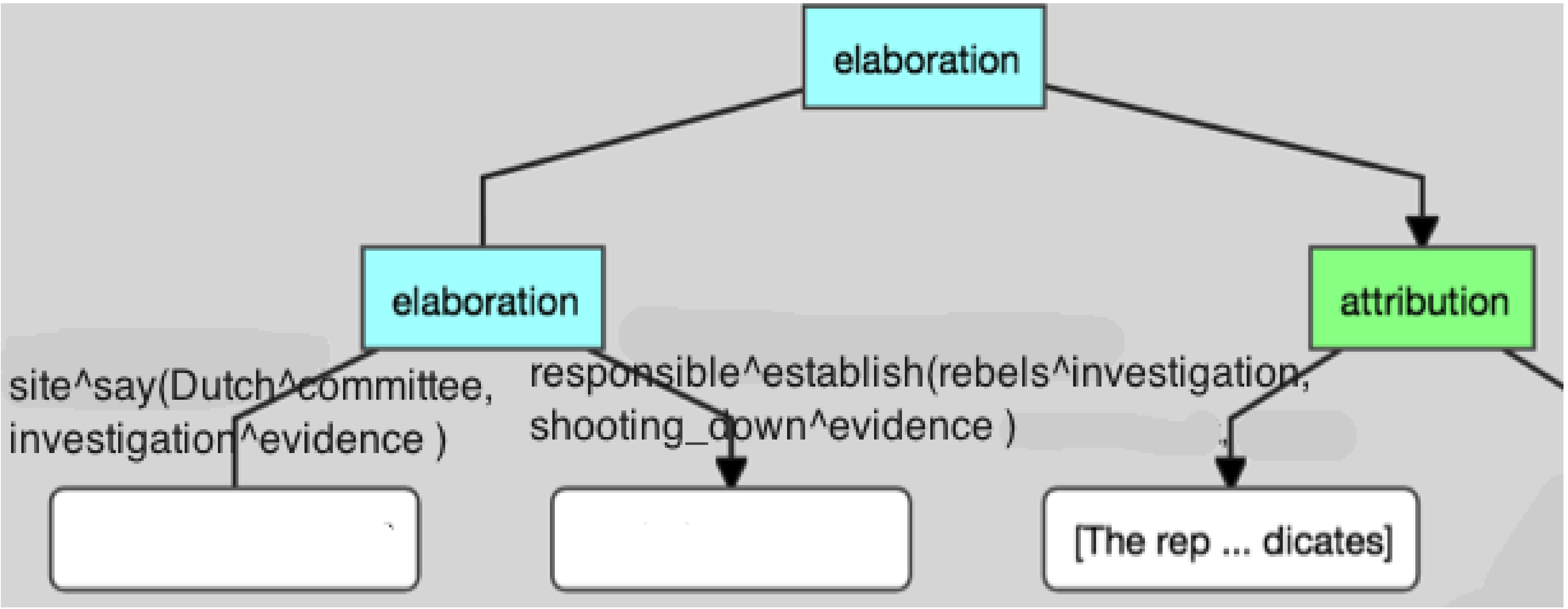
Similarity = number of all common sub- trees

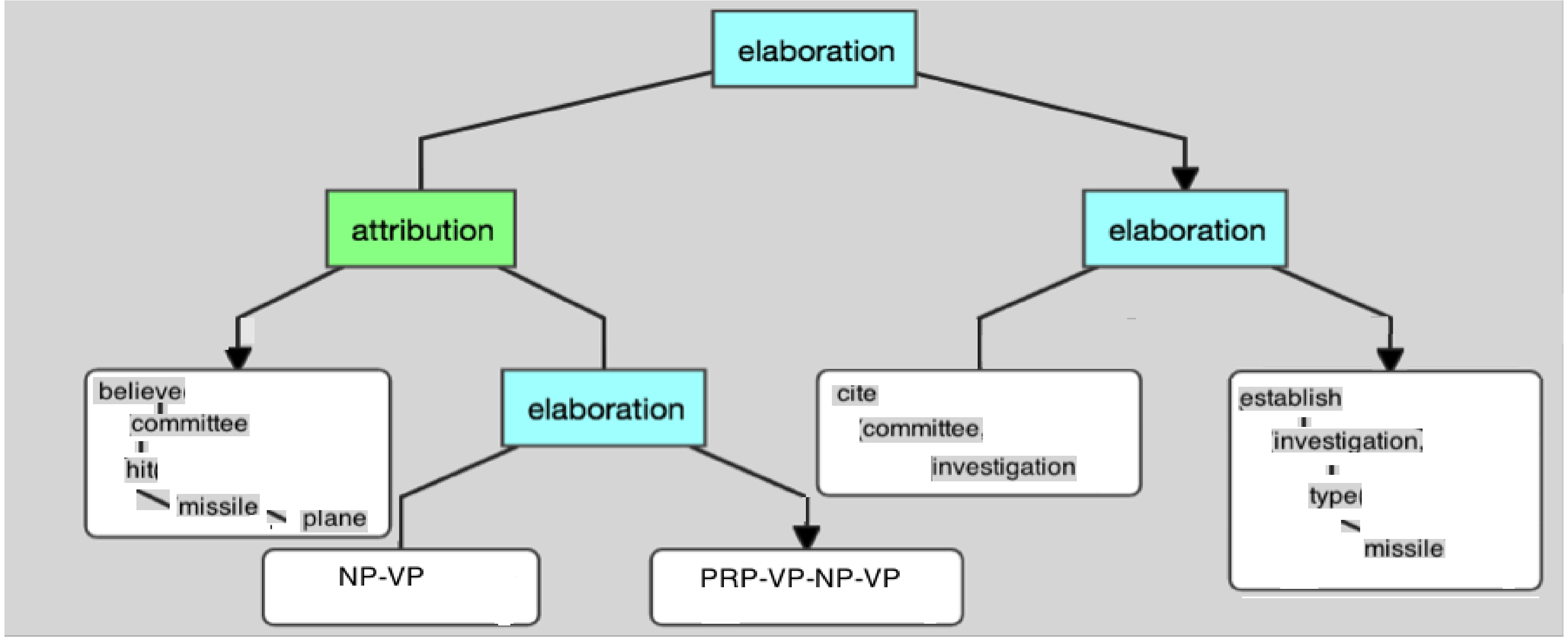
Trees instead of numeric vectors for SVM learning

$$TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2)$$

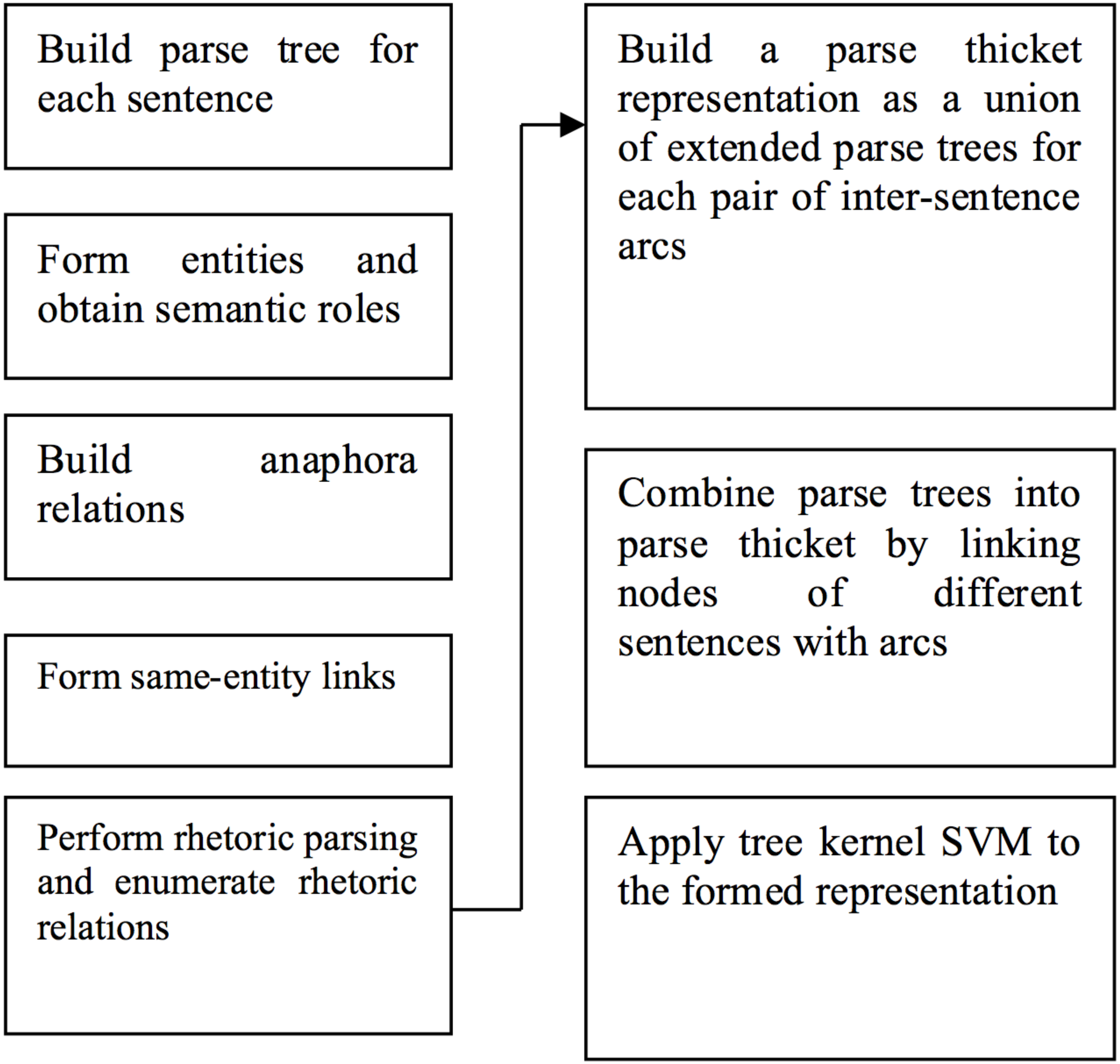
Tree kernel function = number of all common subtrees (Collins et al 2002)

Discourse Tree representation: Maximal Common Subtree vs Tree Kernel





Architecture of learning system



Software components

Stanford NLP Parser, NER, Coreference, Sentiment of (Manning et al 2014, Recasens et al 2013, Lee at al 2013)
VerbNet, JVerbNet (Kipper et al 2008, http://projects.csail.mit.edu/jverbnet/).
OpenNLP.Similarity.parse_thicket

Rhetoric parser of (Surdeanu et al 2015)
Align EDUs for the discourse tree with parse thicket (OpenNLP.Similarity.parse_thicket.matching)
Merge discourse trees with parse thicket.
(OpenNLP.Similarity.parse_thicket.rhetoric_structure)
Obtain 1) Discourse tree with VerbNet signature for CA;
2)Parse thicket with enriched RST relations
Improve text similarity assessment by word2vec model (Mikolov et al 2011, https://deeplearning4j.org/)

Build representation for Thicket Kernel learning
(OpenNLP.Similarity.parse_thicket.kernel_interface)
Build representation for Nearest Neighbor learning
(OpenNLP.Similarity.parse_thicket.jsmlearning)

Apply Thicket Kernel learning (Moschitti 2006, http://disi.unitn.it/moschitti/Tree-Kernel.htm)
Apply Nearest Neighbor learning
(OpenNLP.Similarity.parse_thicket.matching)

Evaluation of Argument and Sentiment Detection

Method / sources	Precision	Recall	F-measure	Precision	Recall	F-measure
	Newspaper opinions			Customer complaints		
Naïve Bag-of-words	63.4	56.7	59.86	52.3	54.2	53.23
WEKA-Naïve Bayes	64.7	57	60.61	56.7	52.6	54.57
SVM TK for RST and CA (full parse trees)	78.8	72.9	75.74	74.6	70.2	72.33
SVM TK for DT	62.4	61.7	62.05	59.3	63.2	61.19
SVM TK for SPDT	81.9	76.3	79.00	75.2	74.6	74.90

This naïve bag-of-words approach is outperformed by the top performing approach (greyed row on the bottom) by 19% for newspapers and 22% for complaints . A Naïve Bayes classifier is improved by 18% and 20% correspondingly.

Data source and method	Preci-sion	Recall	F
Baseline sentiment detector (Stanford NLP)	62.7	68.3	65.38
Hybrid sentiment detector (Stanford NLP + SVM TK for SPDT)	79.3	81.0	80.14
Sentiment detector via SVM TK for *parse thicket*	64.2	66.0	65.09
Sentiment detector via SVM TK for DT	67.5	69.4	68.44
Sentiment detector via SVM TK for SPDT	69.8	68.3	69.04

An improvement of almost 15% is achieved by discovering overall negative sentiment at the paragraph level in case of recognized presence of argumentation. In some of these cases the negative sentiment is implicit and can only be detected indirectly from the discourse structure, where individual words do not indicate negative sentiments.

Evaluation of Q and A Coordination Task

Source / Evaluation setting	Yahoo! Answers			Conversation on Social Networks			Customer complaints			Interviews by journalists		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Types and Counts for rhetoric relations of Q and A	55.2	52.9	54.03	51.5	52.4	51.95	54.2	53.9	54.05	53.0	55.5	54.23
Entity-based alignment of DT of A and A	63.1	57.8	60.33	51.6	58.3	54.70	48.6	57.0	52.45	49.2	57.9	53.21
SVM TK for Parse Trees of individual sentences	66.1	63.8	64.93	69.3	64.4	66.80	46.7	61.9	53.27	78.7	66.8	72.24
SVM TK for RST and CA (full parse trees)	75.8	74.2	74.99	72.7	77.7	75.11	63.5	74.9	68.74	75.7	84.5	79.83
SVM TK for RR-DT	76.5	77	76.75	74.4	71.8	73.07	64.2	69.4	66.69	82.5	69.4	75.40
SVM TK for RR-SPDT	80.3	78.3	79.29	78.6	82.1	80.34	59.5	79.9	68.22	82.7	80.9	81.78
SVM TK for RR-SPDT + sentiment + argumentation features	78.3	76.9	77.59	67.5	69.3	68.38	55.8	65.9	60.44	76.5	74.0	75.21

The highest accuracy is achieved in *journalism and community answers* domain and the lowest in *customer complains* and *social networks*.

The richest source of discourse data (SVM TK for RR-DT) gives the highest classification accuracy, almost the same as the RR similarity-based classification.

The higher is the achieved accuracy having the method fixed, the higher is the level of agreement between Q and A and correspondingly the higher the responder's competence

Results & Conclusions

- Using SVM TK one can differentiate between a broad range of styles of user generated noisy content.
- Each text style and genre has its inherent rhetoric structure that is leveraged and automatically learned.
- When syntactic structure is noisy and some features can be missing, the rhetoric structure with unreliably detected elementary discourse units can still be a reliable indicator of text style.

- Discourse-level technique outperformed traditional keyword-based statistical and/or compositional semantics approaches in three evaluation tasks.
- This improvement is larger for user-generated content in comparison with the professionally written text with proper style and grammar.

Since correlation between text style and text vocabulary is rather low, traditional classification approaches, which only take into account keyword statistics information could lack the accuracy in the complex cases.

README.md

OpenNLP.Similarity Component

It is a project under Apache OpenNLP which subjects results of parsing, part-of-speech tagging and rhetoric parsing to machine learning. It is leveraged in search, content generation & enrichment, chat bots and other text processing domains where relevance assessment task is a key.

What is OpenNLP.Similarity?

OpenNLP.Similarity is an NLP engine which solves a number of text processing and search tasks based on OpenNLP and Stanford NLP parsers. It is designed to be used by a non-linguist software engineer to build linguistically-enabled:

- search engines
- recommendation systems
- dialogue systems
- text analysis and semantic processing engines
- data-loss prevention system
- content & document generation tools
- text writing style, authenticity, sentiment, sensitivity to sharing recognizers
- general-purpose deterministic inductive learner equipped with abductive, deductive and analogical reasoning which also embraces concept learning and tree kernel learning.

OpenNLP similarity provides a series of techniques to support the overall content pipeline, from text collection to cleaning, classification, personalization and distribution. Technology and implementation of content pipeline developed at eBay is described here.