Style and Genre Classification by Means of Deep Textual Parsing

Boris Galitsky[1], Dmitry Ilvovsky[2], Ekaterina Chernyak[2], Sergey O. Kuznetsov[2]

[1] Knowledge Trail Inc., San Jose CA, USA
[2] Higher School of Economics, Moscow, Russia

Outline

1. Problem overview & motivations
2. Domains: styles / genres
3. Representing a paragraph of text
4. Tree kernel learning of paragraphs
5. Evaluation
6. Communicative Discourse Trees & Argument identification problem
7. Conclusions

Problem overview

Classifying a text into abstract classes based on its deep linguistic structure:

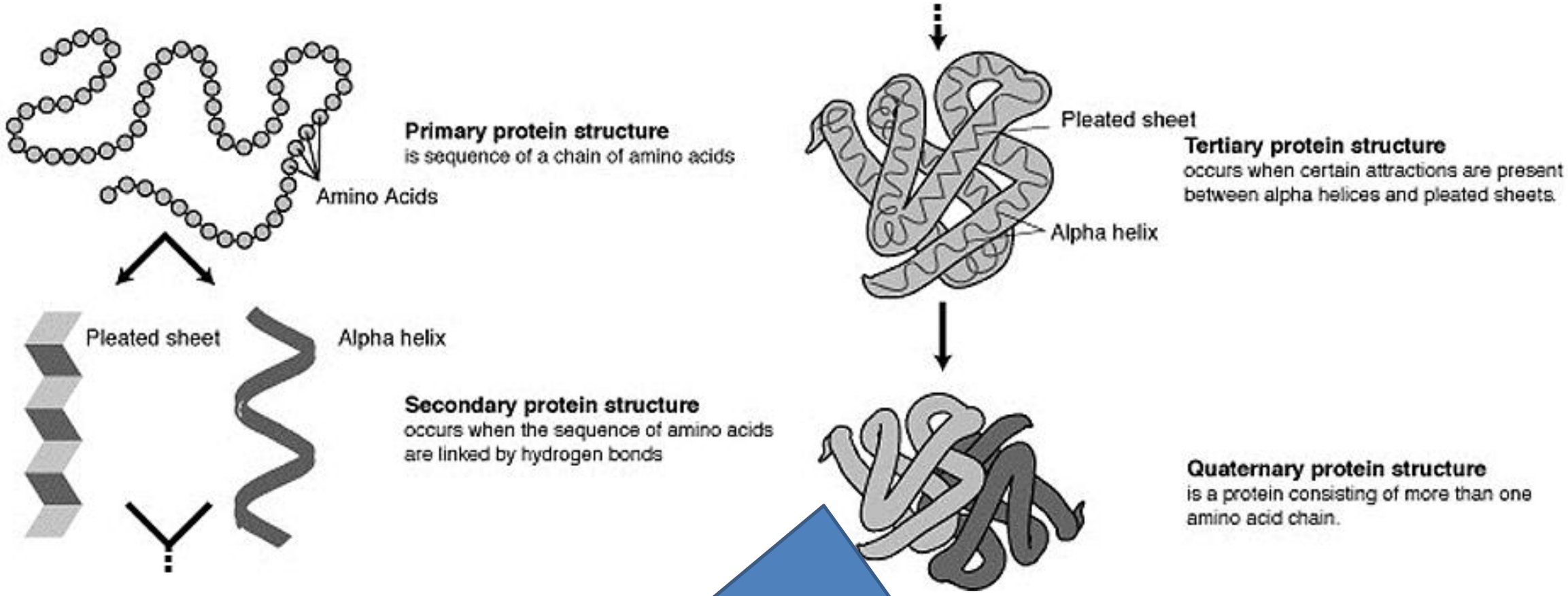
1. Styles
   - 2 classes from industrial task
   - Proprietary Engineering documents vs. Non-proprietary technical documents
2. Genres
   - N classes from academic corpus
   - Corpus based on functional dimensions [Sharoff, 2015]
3. Identifying logical arguments in text
   - Political discussions
   - Customer complaints [Galitsky et al 2008]
   - Disputes / legal documents

Motivation: from keywords to discourse structure

- Just **keywords** - is not enough in many cases
- Presence of **verbs for communicative actions and mental states** - may help to identify a few patterns, but is still insufficient
- **Parse trees** - specific phrases by texts in style or genre but still does not help for systematic exploration of deep linguistic features
- The **discourse structure** including rhetoric relations and communicative actions

We should be able to learn discourse trees

Secondary, Tertiary, Quaternary structure of language



Computational biologists figured out that higher-level structures of protein language help! However, linguists like to talk about higher-level (discourse) structures but neither computationally learn them nor use in applications

Approach

- Build syntactic trees
- Extract discourse information
  - Coreferences
  - Rhetorical relations [Mann et al., 1987]
  - Communicative actions in VerbNet format
- Build additional arcs between syntactic tree nodes of different sentences
- Construct **parse thicket** graph [Galitsky et al 2013]
- Pick **extended trees** from the graph
- Learn on extended trees
- Learn a special case of parse thicket: communicative discourse tree (CDT)

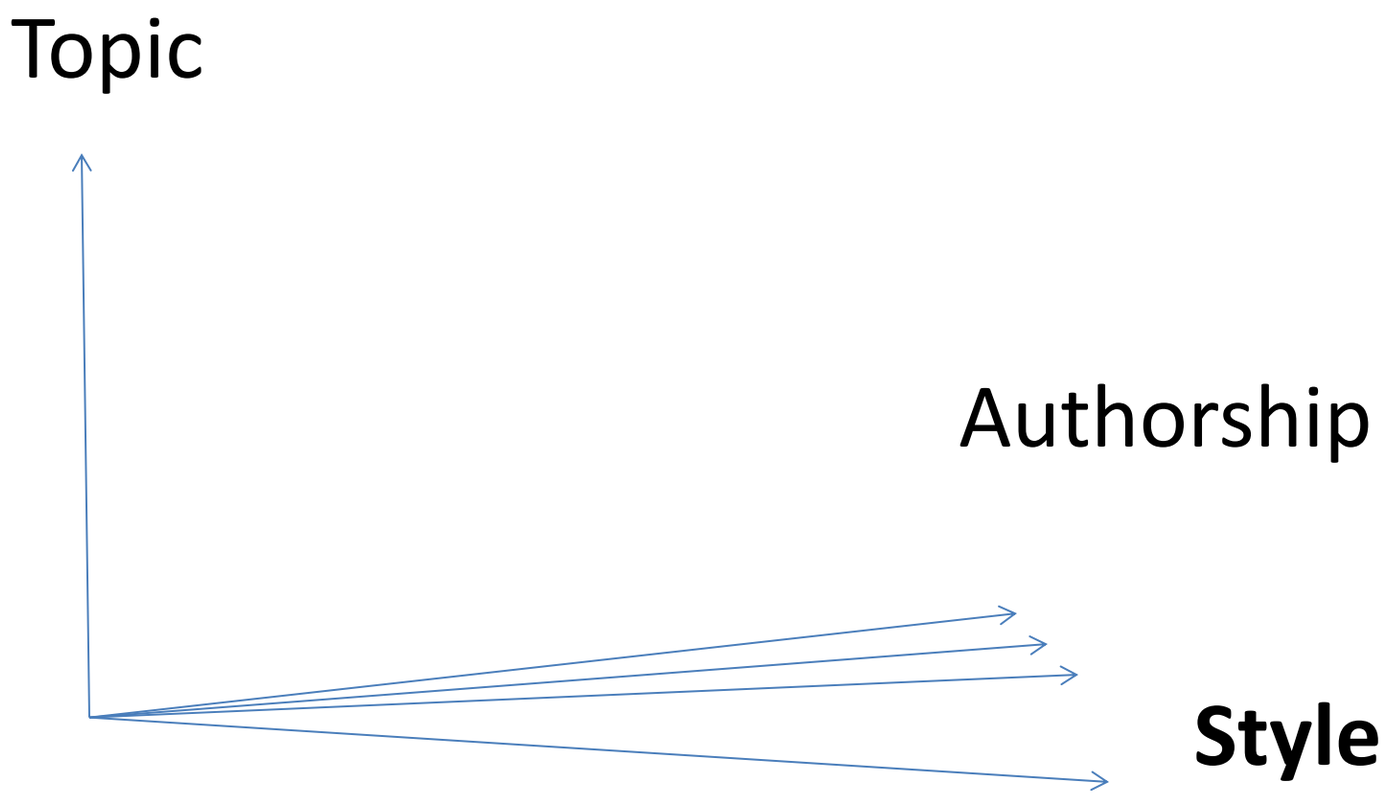
Problem domain 1

- Engineering Design document is a document which contains a thorough and well-structured description of how to build a particular engineering system.
- Not design document : a document which contains meta-level information relatively to the design of engineering system, such as:
  - ✓ how to write design docs manuals,
  - ✓ standards design docs should adhere to,
  - ✓ tutorials on how to improve design documents,
  - ✓ project requirement document,
  - ✓ requirement analysis,
  - ✓ operational requirements,
  - ✓ construction documentation,
  - ✓ project planning and binning,
  - ✓ technical services review, guidelines, manuals, etc.

Problem Domain 2: Genres based onFunctional Text Dimensions

A1	Argum	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', for argumentative blogs, editorials or opinion pieces)
A4	Fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A7	Instruct	To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ)
A8	Hardnews	To what extent does the text appear to be an informative report of events recent at the time of writing?
A9	Legal	To what extent does the text lay down a contract or specify a set of regulations?

Recognizing **document style**, not document topic



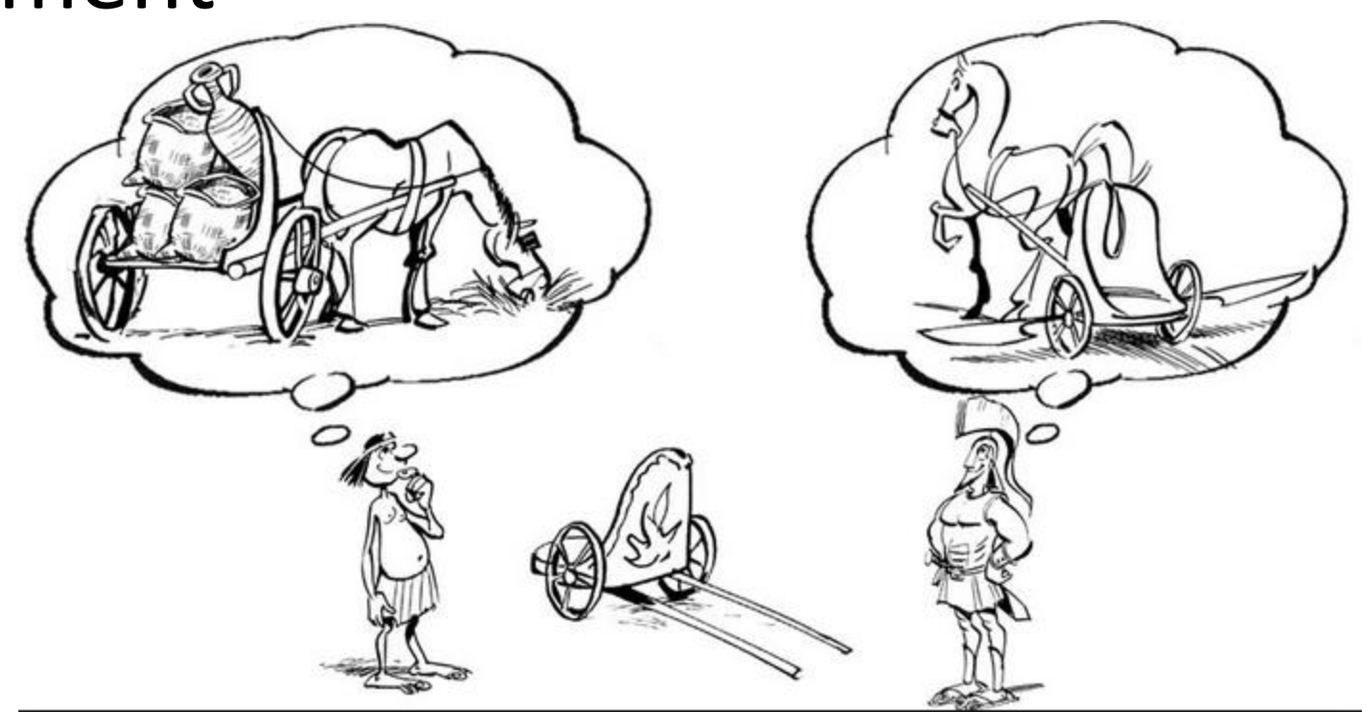
Classification dimensions

- Topicality
- Polarity (sentiment)
- Authorship
- **Document style** (what kind of descriptions is provided for things)

This is necessary for document processing system

Application domains where style recognition is necessary

- Document type (not topic) recognition
- Document flow analysis systems
- Customer relationship management
- Security analysis
- Forensic analysis
- Document quality assessment



Method: Tree Kernels

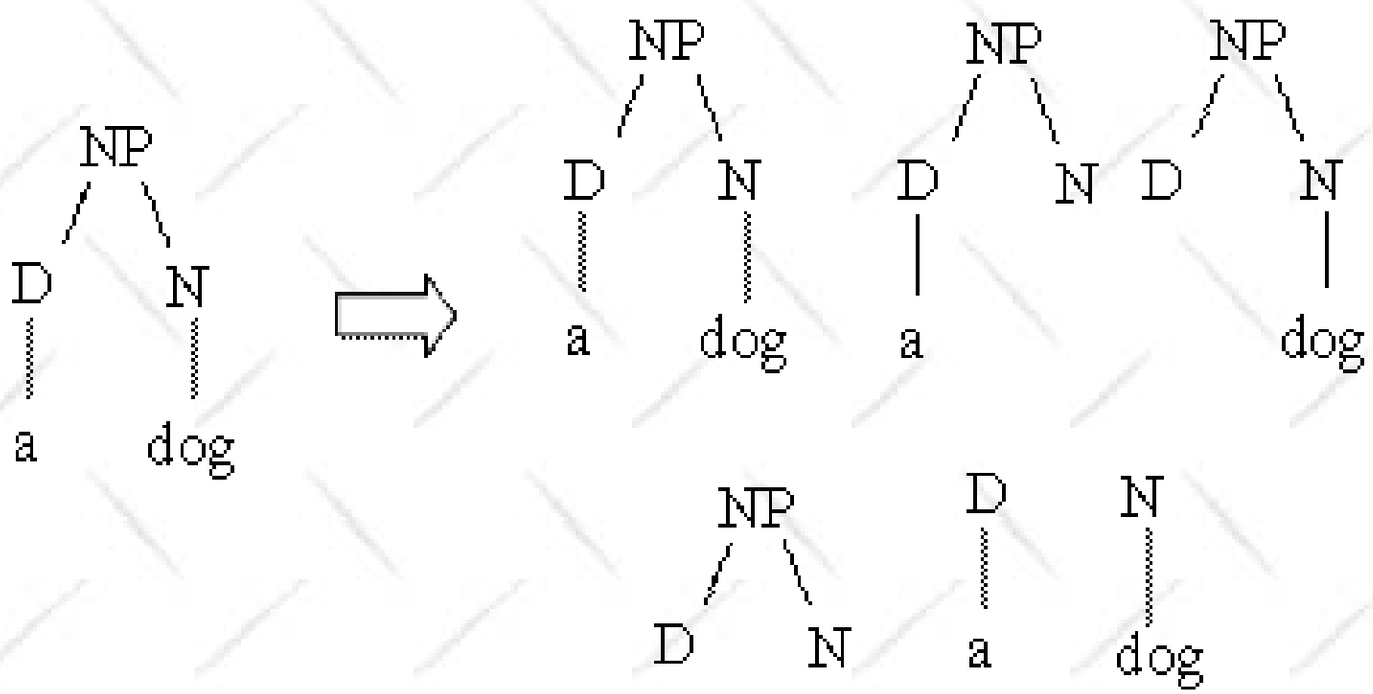
Tree Kernels

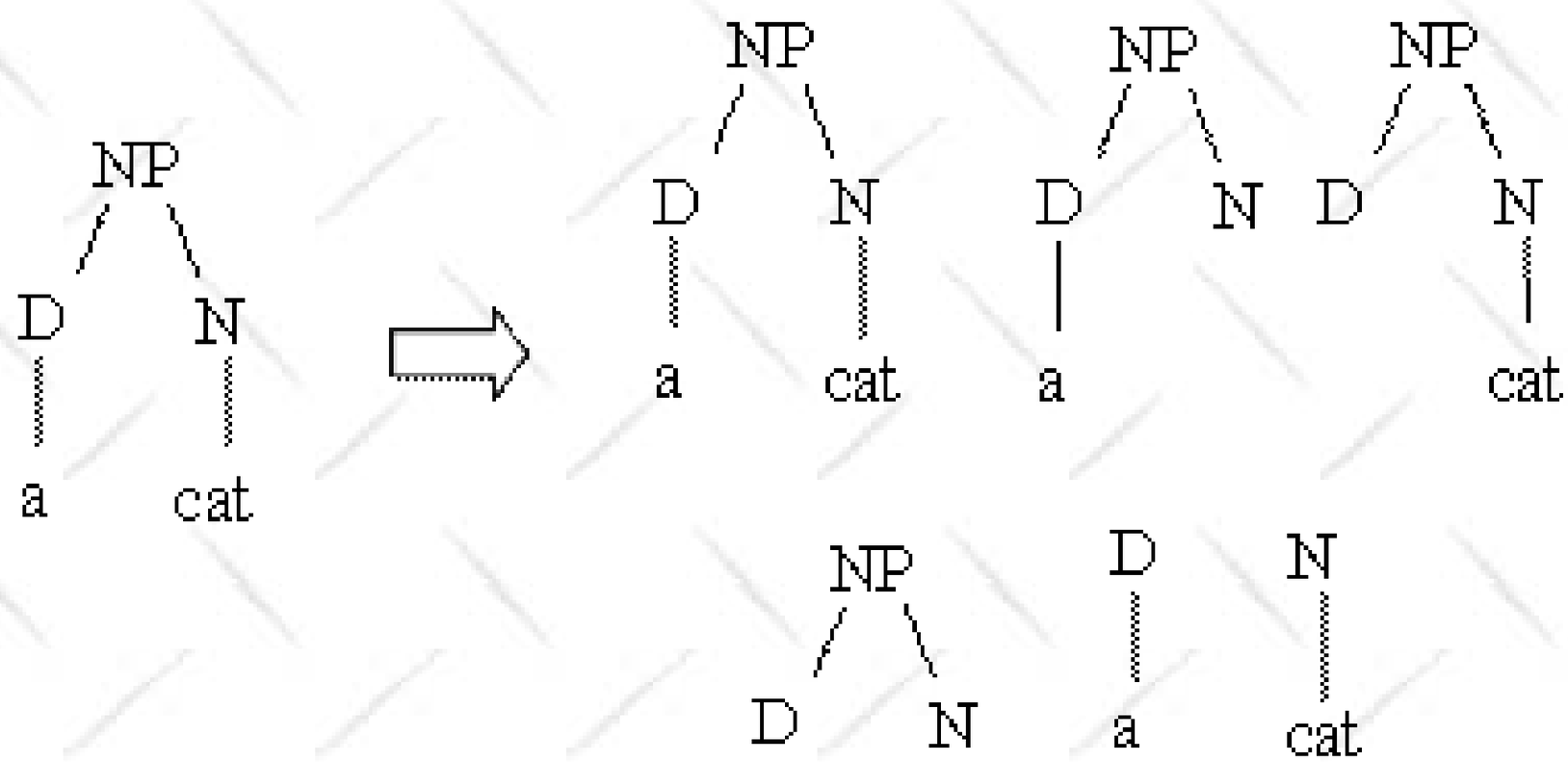
What are they and why use them for classification?

We need to measure distances between texts:

- Traditionally – using statistic of keywords
- More accurate – using statistics of all sub-parse trees

Measuring similarity between phrases





SVM Tree Kernel learning

- **Trees** instead of numeric vectors

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

- Tree kernel function = number of all common subtrees [Collins et al 2002]

Applications of Tree Kernels

Convolution tree kernel (Collins and Duffy, 2002) defines a feature space consisting of all subtree types of parse trees and counts the number of common subtrees as the syntactic similarity between two parse trees. They have found a number of applications in a series of NLP tasks, including

- search (Moschitti 2006);
- syntactic parsing re-ranking;
- relation extraction (Zelenko et al 2003);
- named entity recognition (Cumby & Roth 2003) ;
- Semantic Role Labeling / relation extraction (Zhang et al 2006);
- pronoun resolution (Yang et al., 2006);
- question classification and machine translation (Zhang and Li 2009).

Limitation of sentence-level tree kernels

Phrases important for classification can be distributed through different sentences

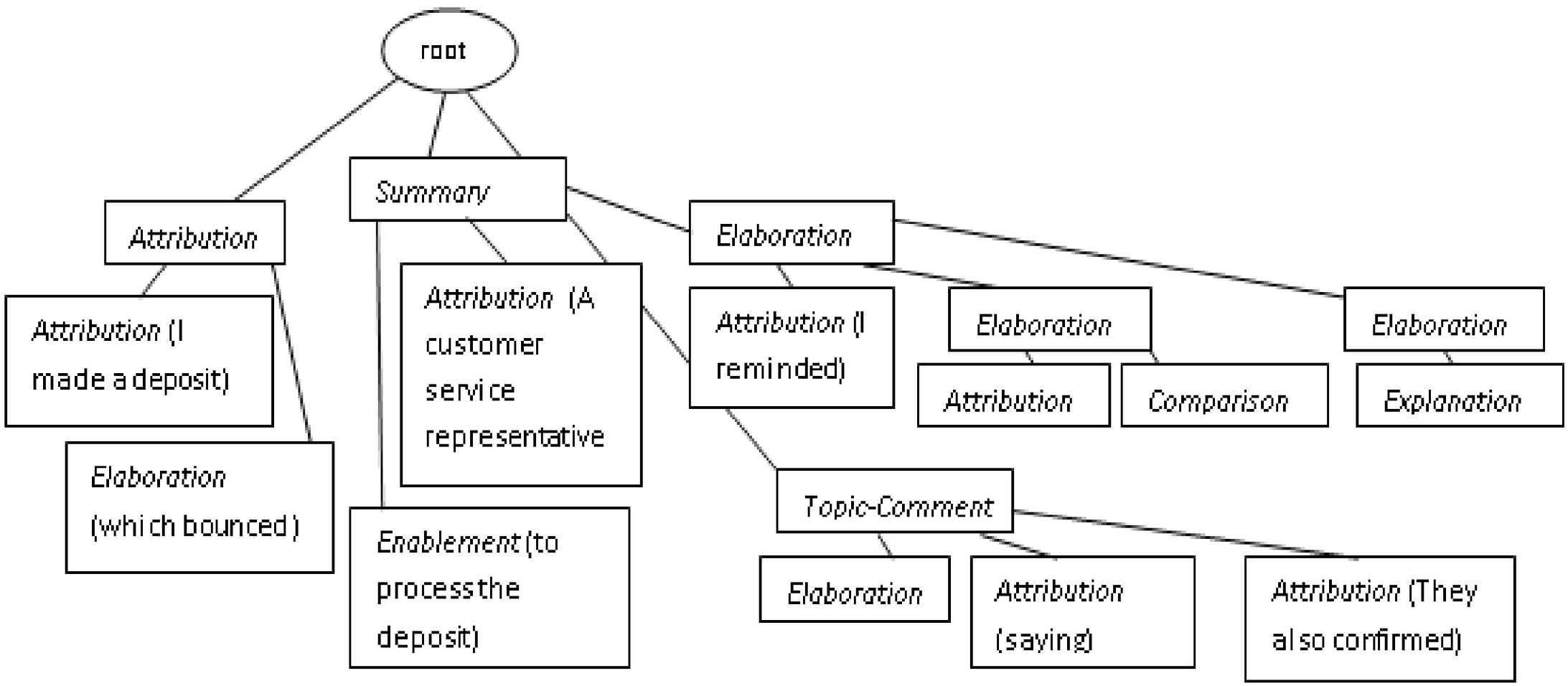
So we want to combine/merge parse trees to make sure we cover the phrase of interest.

This document describes how the back end processor can be designed. Its requirements are enumerated below.

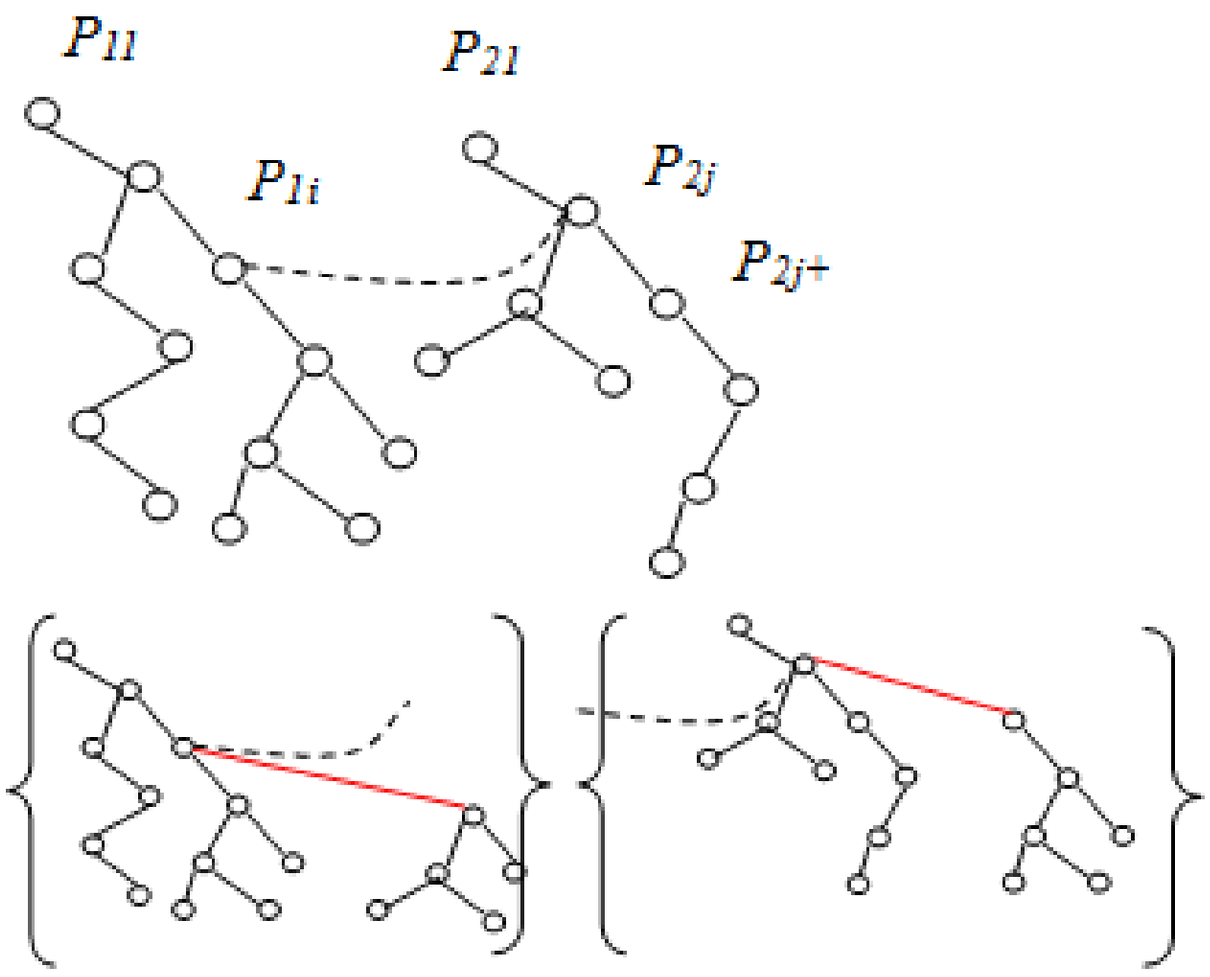
From the first sentence, it looks like we got the design doc. To process the second sentence, we need to disambiguate 'its'. As a result, we conclude from the second sentence that it is a requirement doc

Rhetoric structure

Discourse tree for a paragraph of text

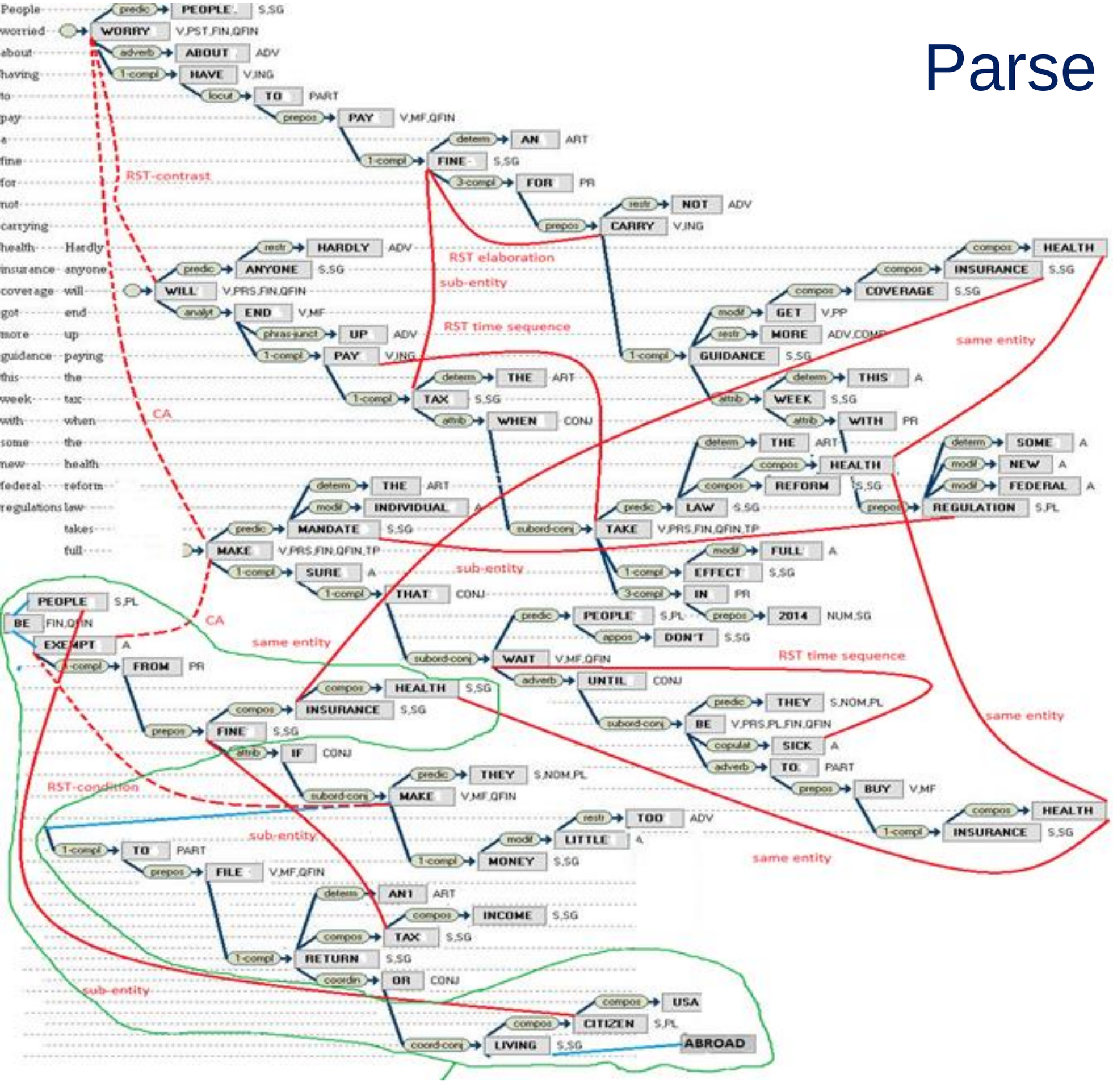


Merging parse trees, once inter-sentence link is found

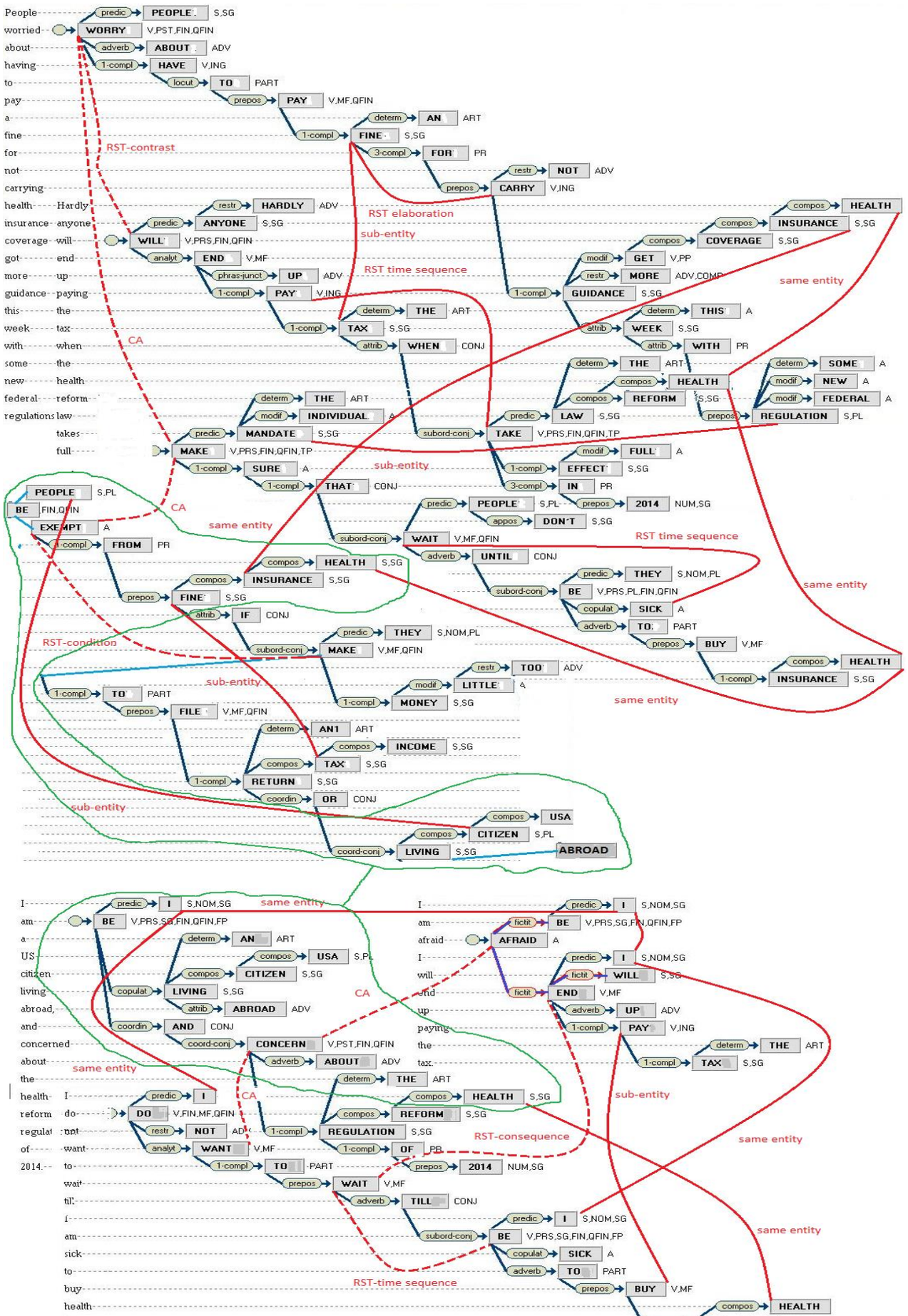


Paragraph Level Kernel = Sum of kernels for all extended tree pairs

Parse Thicket



Measuring similarity between two texts using Parse Thickets



Evaluation 1 & 2

Evaluation setup. Styles

- Set of 1200 documents, auto web mined using keywords "design document AND (java OR engineering OR hardware OR software …)
- Total 30 search keywords (domains)
- 90% of non-design engineering documents of the classes we want to exclude (meta-documents) and 10% of genuine design documents
- 5 sub-sets for training/evaluation portions
- Evaluation results were assessed by quality assurance personnel
- Relevant dataset for benchmarking is unavailable. Evaluation set is submitted to UCI Machine Learning repository

Results. Styles

Method	Precision,%	Recall,%	F-measure
*«Nearest neighbor» (based on TF*IDF)*	53.9	62	57.67+-0.62
Naïve Bayesian	55.3	59.7	57.42+-0.84
Tree kernel SVM – regular parse trees	71.4	76.9	74.05+-0.55
Tree kernel SVM – extended parse trees for anaphora	77.8	81.4	79.56+-0.70
Tree kernel SVM – extended parse trees for RST	80.1	80.5	80+-1.03
Tree kernel SVM – extended parse trees for both anaphora and RST	**83.3**	**83.6**	**83.45+-0.78**

Evaluation. Genres

- Dataset
  - Genre dataset collected by [Sharoff, 2013]
  - Each genre - linear combination of Functional Text Dimensions (FTD)
  - Human-annotated: 0.5, 1, 2 of each FTD
  - 17 FTD
- Evaluation
  - 7 genres: fiction stories, Ted, news etc.
  - 9 FTD
  - Pairwise genre classification
  - SVM on extended parse trees for both anaphora and RST

Evaluation. Examples of genres

Genre example	A1	A3	A4	A5	A6	A7	A8	A9	A11
[ted] Emotional political speech	1	0	0	1	0	0	0	0	2
[fict] Fiction	0	1	2	1	0.5	0	0	0	1
[news] News article	0.5	1	0	1	1	0	2	0	0.5
[synd] Article on political topic	2	0	0	0	0	0.5	0.5	0	0
[un] UN reports	1	0	0	0	0	0.5	0	2	0
[tele] Hardware manuals	0	0	0	0	0	2	0	0	0
[tells] Software manuals	0	0	0	0	0	0.5	0	0	1

Results. Genres

Classes	Recall	Precision	F-measure
Fict vs News	98.11	95.55	96.81
Ted vs Synd	99.49	98.94	99.21
Un vs News	98.70	94.93	96.78
Tele vs Tells	96.69	90.76	93.63
Fict vs Ted	97.12	93.74	95.40
Fict vs Synd	95.21	94.23	94.72
Fict vs Un	94.90	95.71	95.30
Fict vs Tele	97.25	94.90	96.06
Fict vs Tells	96.51	95.61	96.06
News vs Ted	96.85	93.85	95.33
News vs Synd	97.28	96.19	96.73
News vs Tele	96.31	94.27	95.28
News vs Tells	98.28	96.15	97.20
Ted vs Un	96.45	97.03	96.74
Ted vs Tele	97.28	96.62	96.95

Analysis

- **17 % increase in style classification** by using tree kernels
- **Further 8% increase** by using extended parse trees
- **Rhetoric relation** source has the strongest contribution
- **Adding new** sources of information increases performance

Domain3: Identifying Logical Argument in Text

What is required to represent an argument pattern in text?

We build an RST representation of the arguments and observe if a discourse tree is capable of indicating whether a paragraph communicates both:

- a claim, and
- an argumentation that backs it up.

We then explore what needs to be added to a discourse tree (DT) so that it is possible to judge if it expresses an argumentation pattern or not.

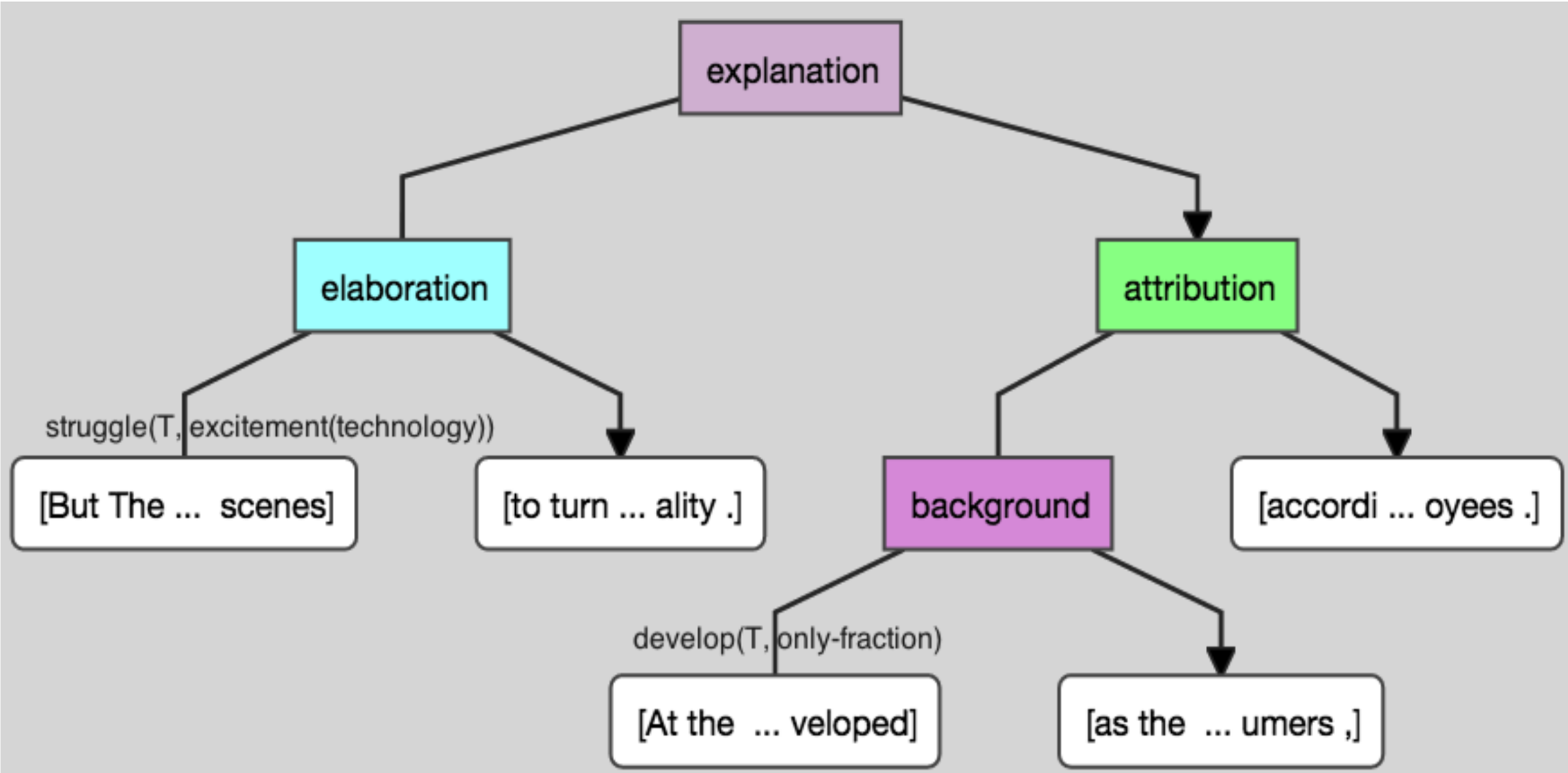
Our example

We present a controversial article published in Wall Street Journal about Theranos, a company providing healthcare services, and the company rebuttal (Theranos 2015).

"Since October [2015], the Wall Street Journal has published a series of anonymously sourced accusations that inaccurately portray Theranos. Now, in its latest story ("U.S. Probes Theranos Complaints," Dec. 20), the Journal once again is relying on anonymous sources, this time reporting two undisclosed and unconfirmed complaints that allegedly were filed with the Centers for Medicare and Medicaid Services (CMS) and U.S. Food and Drug Administration (FDA)."

A claim and its CDT

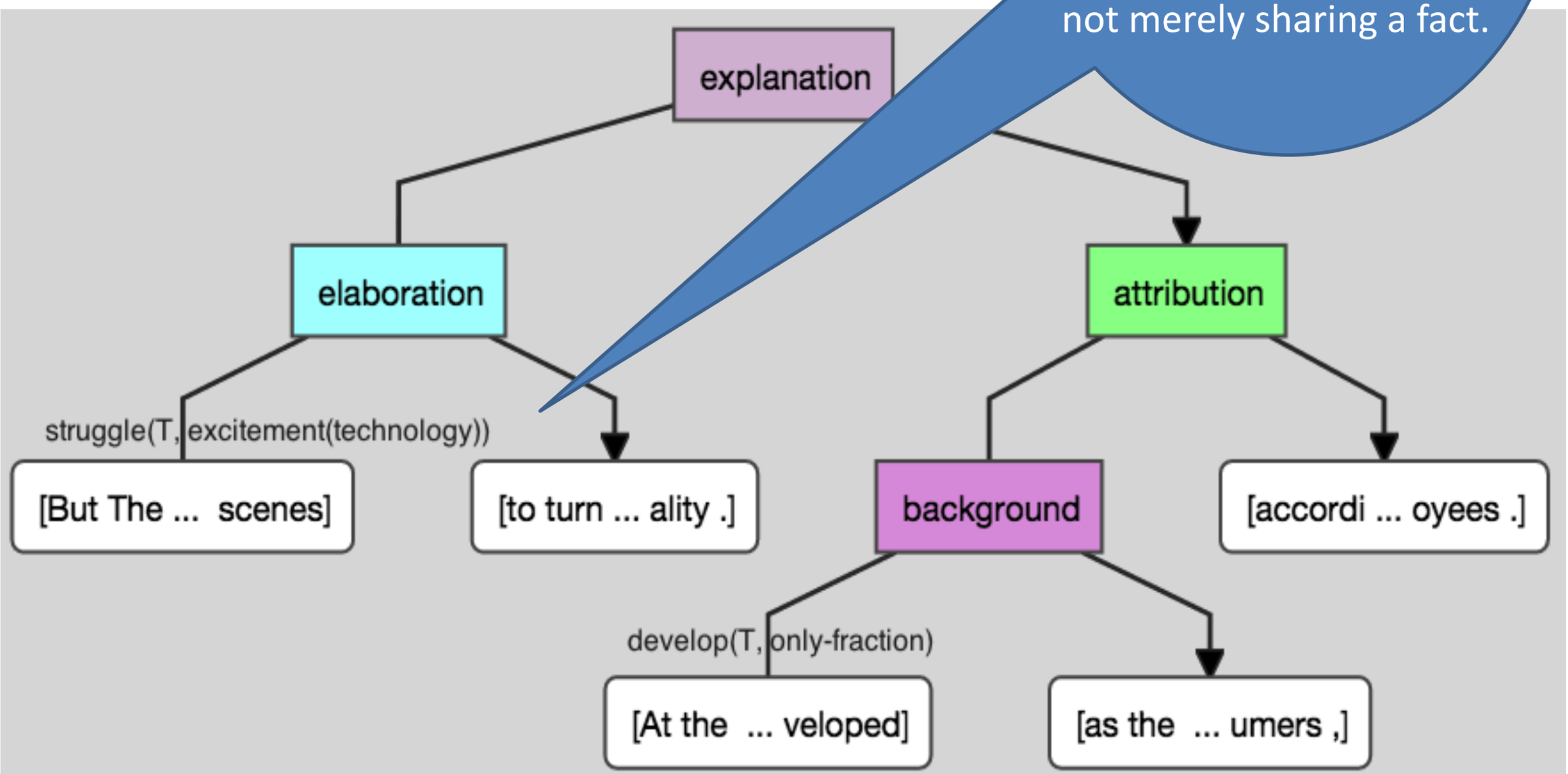
"…But Theranos has struggled behind the scenes to turn the excitement over its technology into reality. At the end of 2014, the lab instrument developed as the linchpin of its strategy handled just a small fraction of the tests then sold to consumers, according to four former employees."



A claim and its CDT

"…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."

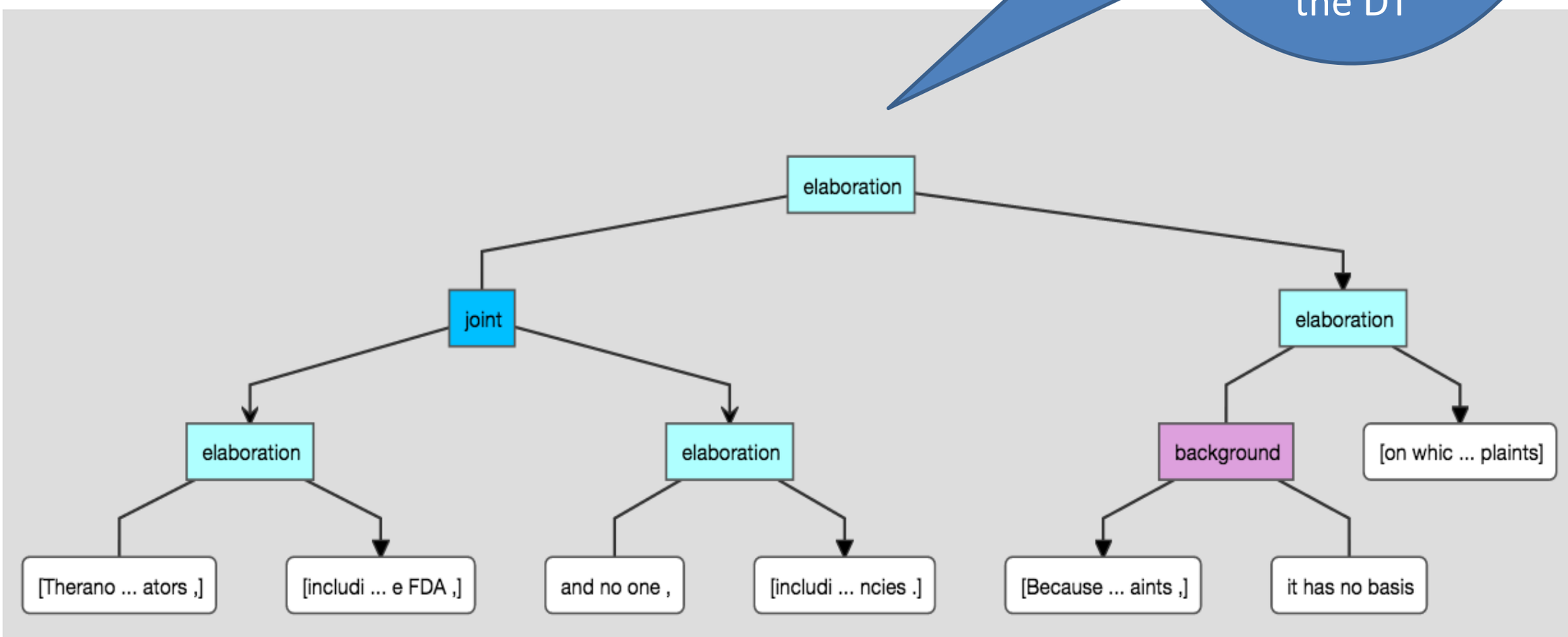
When arbitrary communicative actions are attached to DT as labels of its terminal arcs, it becomes clear that the author is trying to bring her point across and not merely sharing a fact.



A claim and its DT

"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."

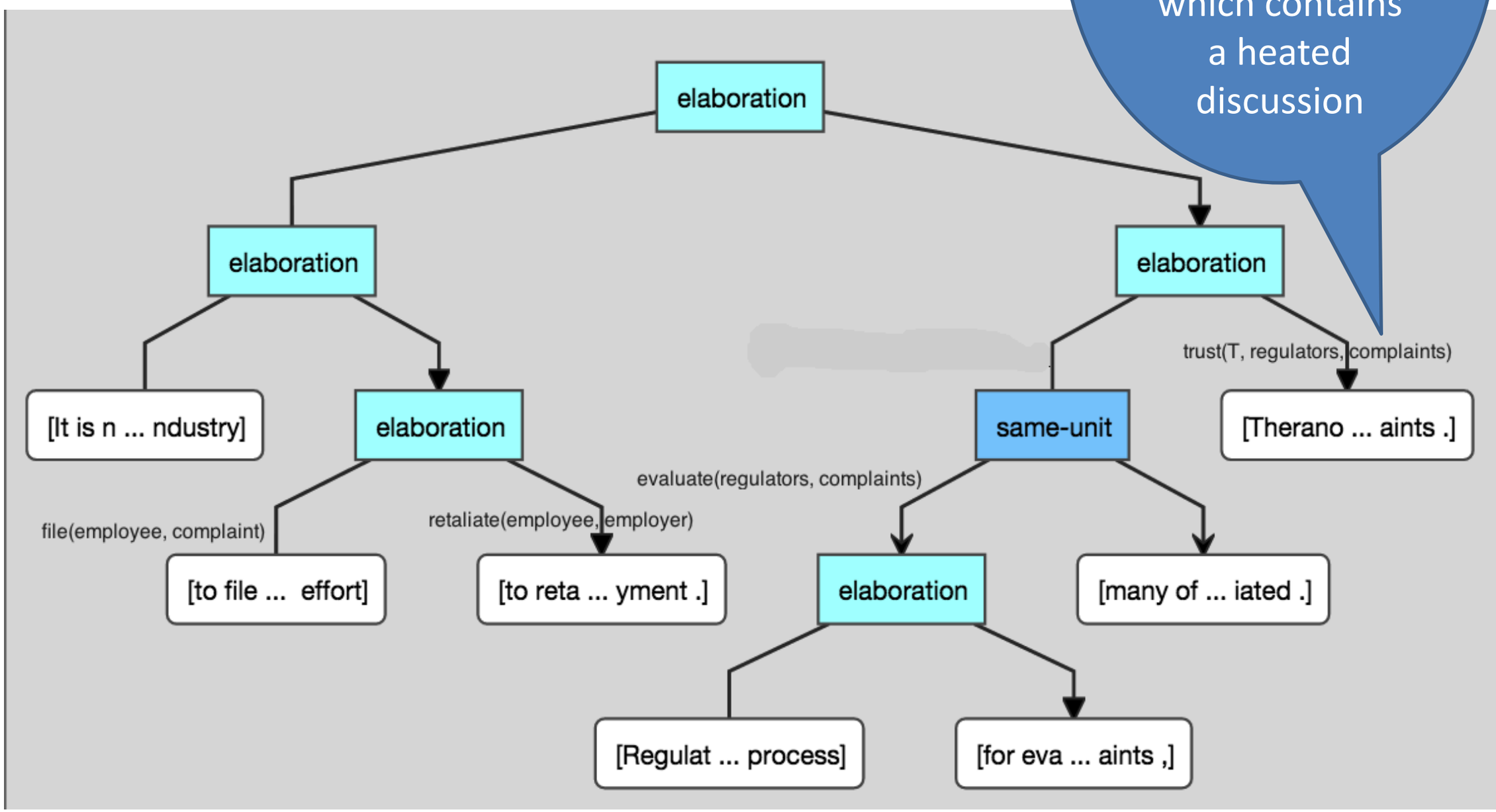
Theranos attempts to rebuke the claim of WSJ, but without communicative actions it is unclear from the DT



CDT for an attempt by Theranos to acquit itself

"It is not unusual for disgruntled and terminated employees in the heavily regulated health care industry to file complaints in an effort to retaliate against employers for termination of employment. Regulatory agencies have a process for evaluating complaints, many of which are not substantiated. Theranos trusts its regulators to properly investigate any complaints."

Communicative actions as labels for rhetoric relations helps to identify a text which contains a heated discussion

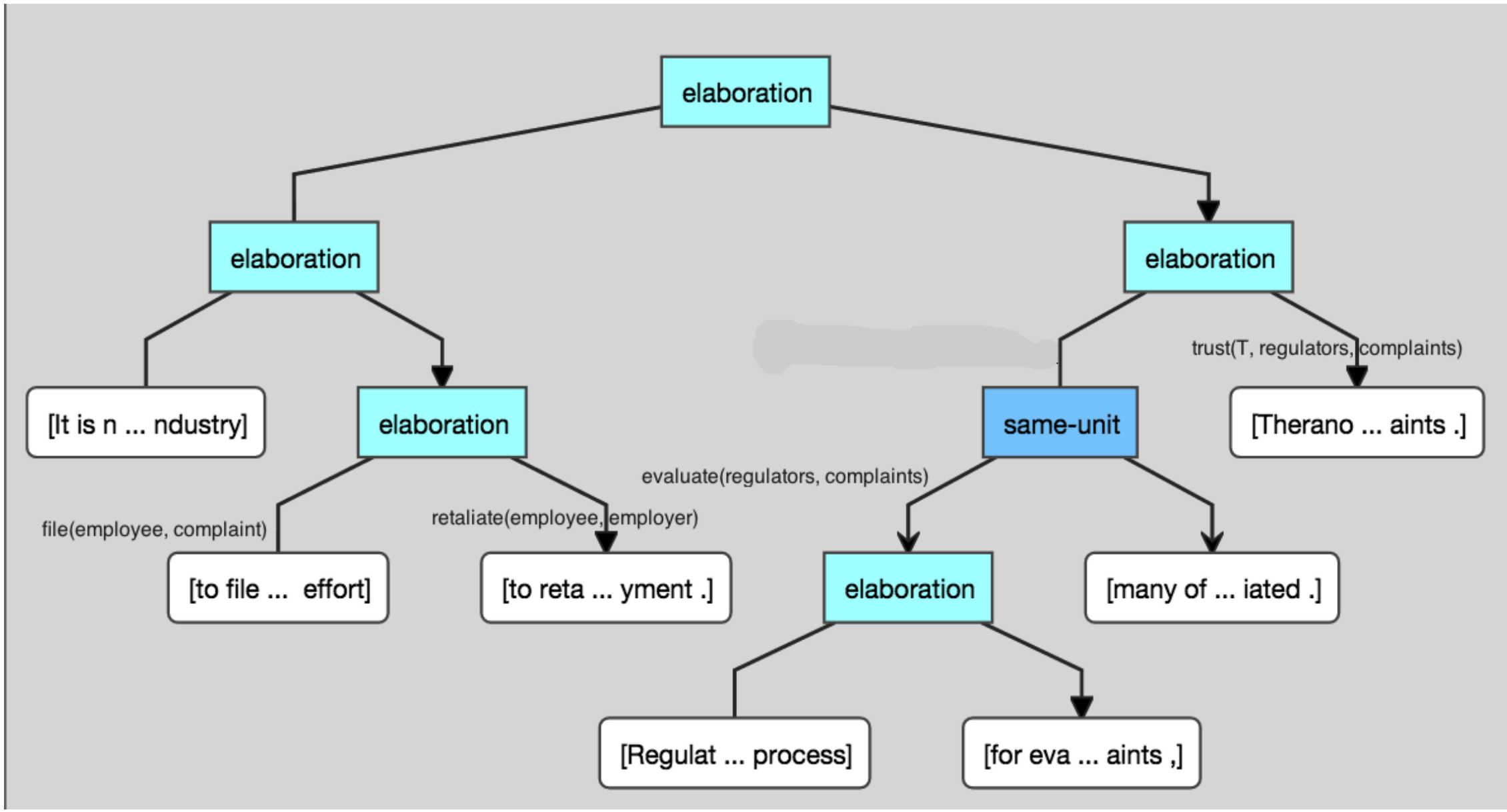


RST + CA => identification of logical argumentation

- To show the structure of arguments, discourse relations are necessary but insufficient, and communicative actions, or speech acts are necessary but insufficient as well.
- We need to know the discourse structure of interactions between agents, and what kind of interactions they are.
- We don’t need to know domain of interaction (here, health), the subjects of these interaction (the company, the journal, the agencies), what are the entities, but we need to take into account mental, domain-independent relations between them.

Another CDT for an attempt to make even stronger counter-claim

"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."



Extracting communicative actions from text: VerbNet entry

Let us consider the verb *amuse*.

There is a cluster of similar verbs that have a similar structure of arguments (semantic roles) such as *amaze, anger, arouse, disturb, irritate,* and other.

The roles of the arguments of these communicative actions are as follows:

- Experiencer (usually, an animate entity)
- Stimulus
- Result

Extracting communicative actions from text: VerbNet Frames

NP V NP
Example: "*The teacher amused the children*."
Syntax:Stimulus V Experiencer
Clause:
amuse(Stimulus, E, Emotion, Experiencer):-
cause(Stimulus, E),
emotional_state(result(E), Emotion, Experiencer).

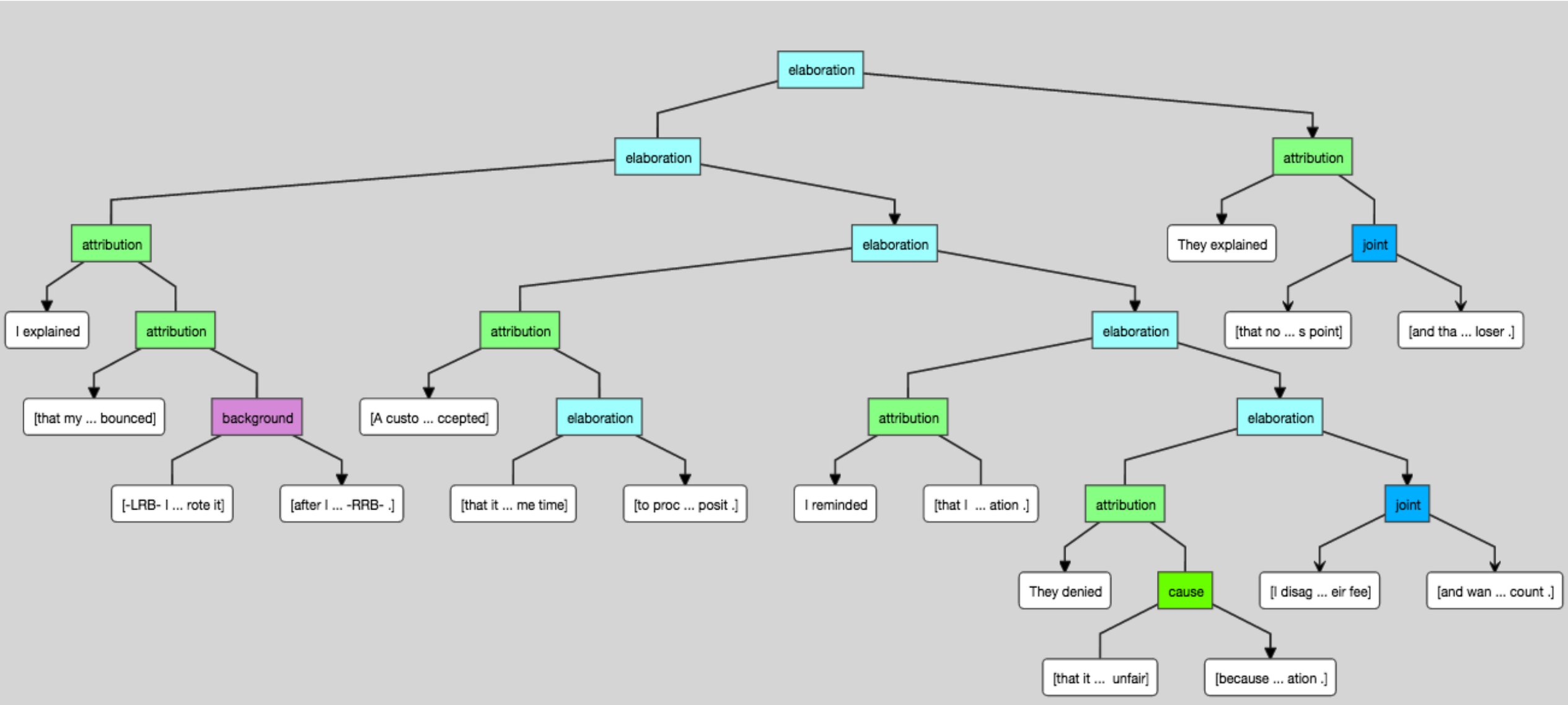
NP V ADV-Middle
Example: "*Small children amuse quickly*."
Syntax:Experiencer V ADV
Clause:
amuse(Experiencer, Prop):-
property(Experiencer, Prop), adv(Prop).

For this example, the information for the class of verbs *amuse is at* http://verbs.colorado.edu/verb-index/vn/amuse-31.1.php#amuse-31.1

NP V NP-PRO-ARB
example "The teacher amused."
syntax Stimulus V
amuse(Stimulus, E, Emotion, Experiencer):-
cause(Stimulus, E),
emotional_state(result(E), Emotion, Experiencer).

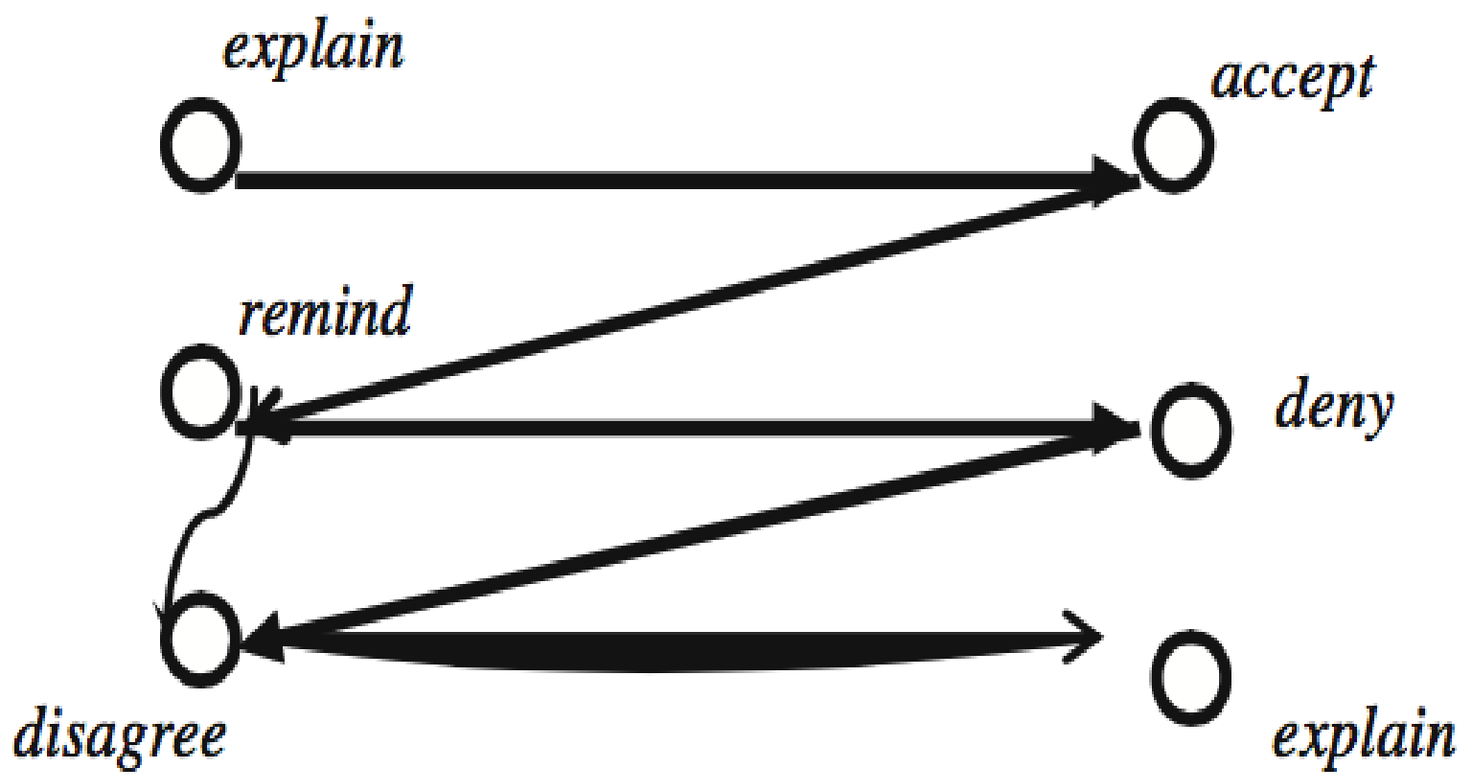
Argumentation scenario and its DT

*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.*



Argumentation scenario and its graph representation

*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.*



Extracting communicative actions from text: clusters

Verbs with Predicative Complements
Appoint, characterize, dub, declare, conjecture, masquerade, orphan, captain, consider, classify.
Verbs of Perception
See, sight, peer.
Psych-Verbs (Verbs of Psychological State)
Amuse, admire, marvel, appeal.
Verbs of Desire
Want, long.
Judgment Verbs
Judgment.
Verbs of Assessment
Assess, estimate.
Verbs of Searching
Hunt, search, stalk, investigate, rummage, ferret.
Verbs of Social Interaction
Correspond, marry, meet, battle.
Verbs of Communication
Transfer(message), inquire, interrogate, tell, manner(speaking), talk, chat, say, complain, advise, confess, lecture, overstate, promise.
Avoid Verbs
Avoid.
Measure Verbs
Register, cost, fit, price, bill.
Aspectual Verbs
Begin, complete, continue, stop, establish, sustain.

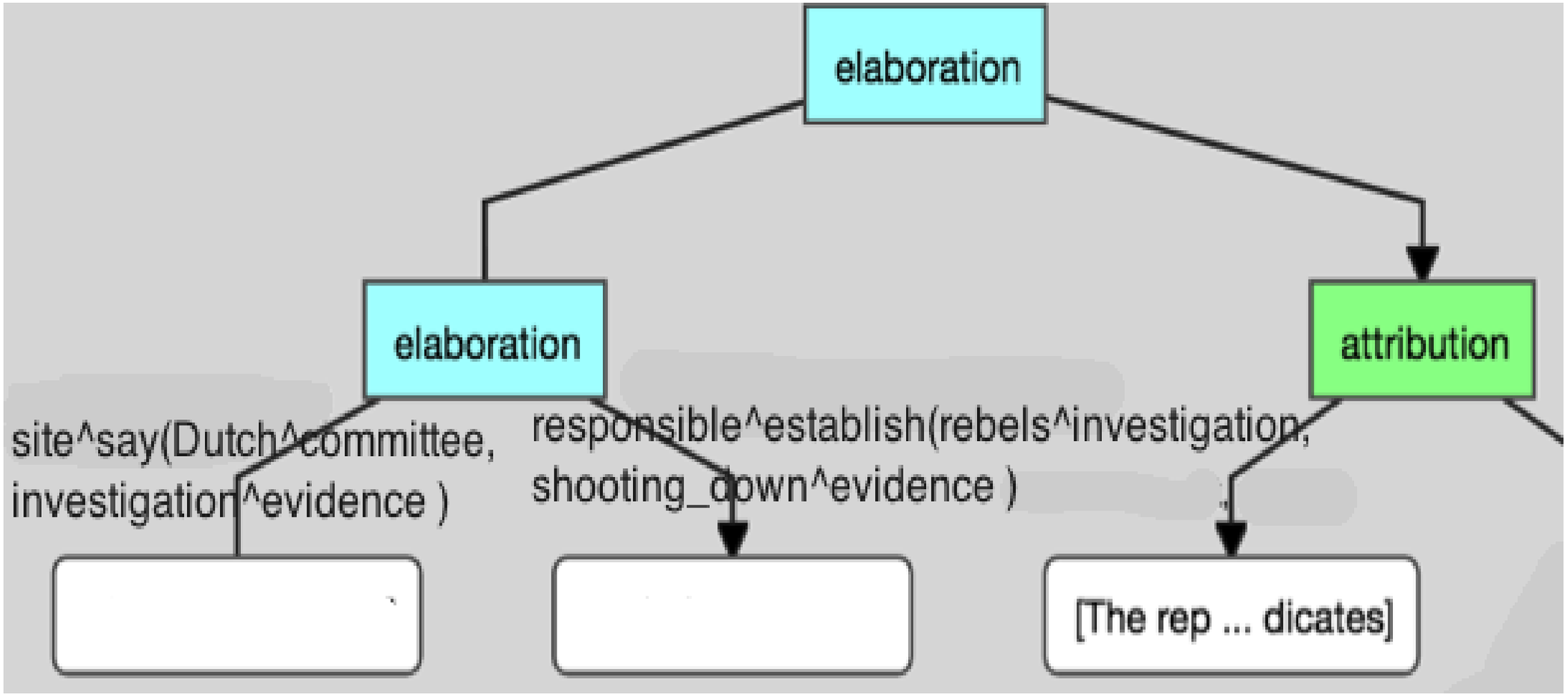
Statistical vs inductive learning

To compute similarity between abstract structures, two approaches are employed:

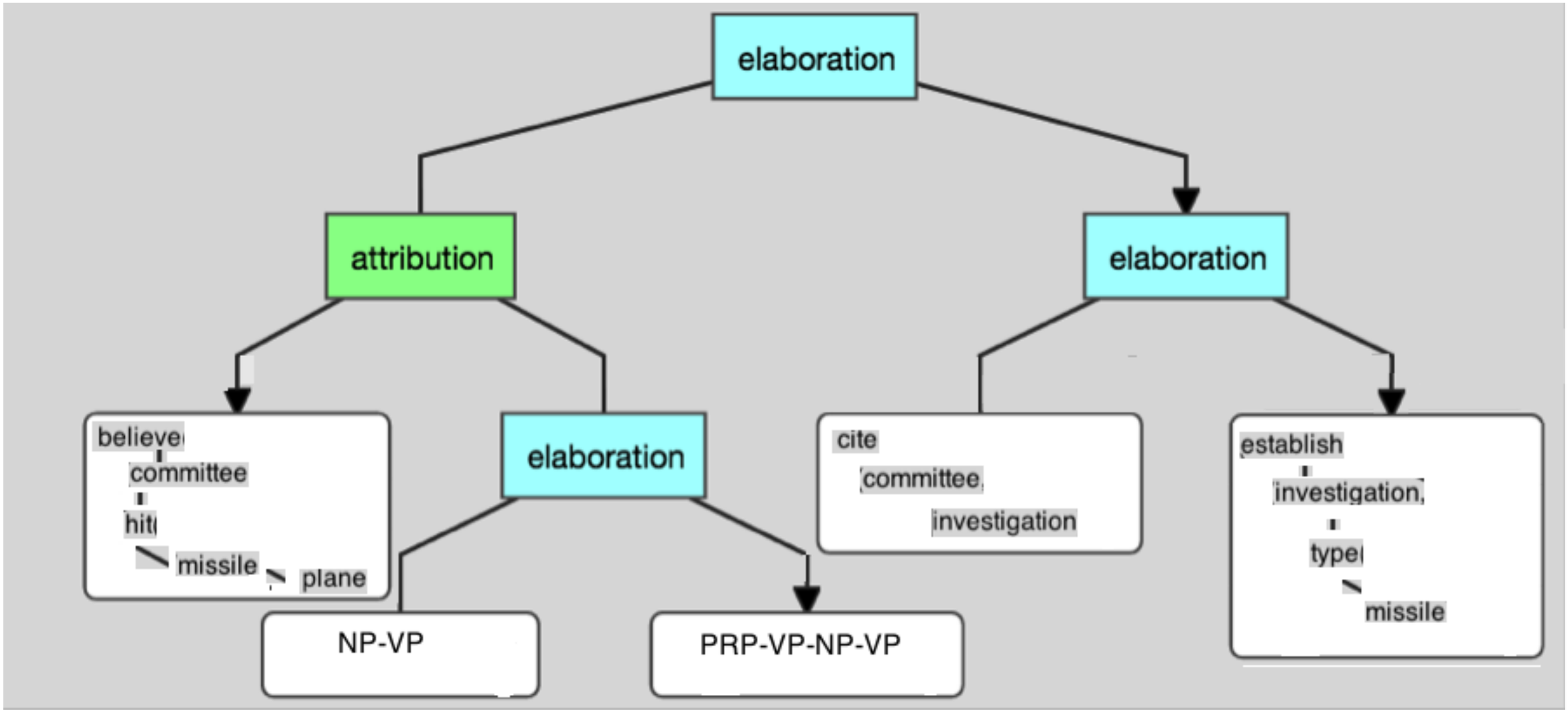
- represent these structures in a numerical space, and express similarity as a number. This is a statistical learning approach
- use a structural representation, without numerical space, such as trees and graphs, and express similarity as a maximal common sub-structure. We refer to such operation as *generalization*. This is an inductive learning approach.

To conduct feature engineering, we will compare both these approaches for text classification. The representation machinery and learning settings are different, but the classification accuracies can be compared.

Similarity between CDTs: maximal common sub-CDTs with generalized labels



Numerical similarity between CDTs: representation for kernel model



Conclusions and future work

- Extended tree kernel performs classification where “typical” text classification algorithms do not perform well => **structure rules**
- Extended tree kernel significantly outperforms regular ones => **discourse matters**
- “1 vs others scheme”, comparison with methods based on character n-gram features for genre classification => **work to be done**