Validating Correctness of Textual Explanation with Complete Discourse trees

Boris Galitsky
Oracle Corp Redwood Shores CA USA
And
Dmitry Ilvovsky
National Research University Higher School of Economics, Moscow, Russia

Abstract

We explore how to validate the soundness of textual explanations in a domain-independent manner. We further assess how people perceive explanations of their opponents and what are the factors determining whether explanations are acceptable or not. We discover that what we call a *complete discourse tree* (complete DT) determines the acceptability of explanation. A complete DT is a sum of a traditional DT for a paragraph of actual text and an imaginary DT for a text about entities used but not explicitly defined in the actual text.

1 Introduction

Providing explanations of decisions for human users, and understanding how human agents explain their decisions, are important features of intelligent decision making and decision support systems. A number of complex forms of human behavior is associated with attempts to provide acceptable and convincing explanations. In this paper, we propose a computational framework for assessing soundness of explanations and explore how such soundness is correlated with discourse-level analysis.

Importance of the explanation-aware computing has been demonstrated in multiple studies and systems. Also, (Walton, 2007) argued that the older model of explanations as a chain of inferences with a pragmatic and communicative model that structures an explanation as a dialog exchange. The field of explanation-aware computing is now actively contributing to such areas as legal reasoning, natural language processing and also multi-agent systems (Dunne and Bench-Capon, 2006). It has been shown (Walton, 2008) how the argumentation methodology implements the concept of explanation by transforming an example of an explanation into a formal dialog structure. Galitsky (2008) differentiated between explaining as a chain of inference of facts mentioned in dialogue, and meta-explaining as dealing with formal dialog structure represented as a graph. Both levels of explanations are implemented as argumentation: explanation operates with individual claims communicated in a dialogue, and meta-explanation relies on the overall argumentation structure of scenarios.

In this paper we explore how good explanation in text can be computationally differentiated from bad explanation. Intuitively, a *good* explanation convinces the addressee that a communicated claim is right, and it involves valid argumentation patterns, logical, complete and thorough. A bad explanation is unconvincing, detached from the beliefs of the addressee, includes flawed argumentation patterns and omits necessary entities. In this work we differentiate between good and bad explanation based on a *human response* to such explanation. Whereas users are satisfied with good explanation by a system or a human, bad explanations usually lead to dissatisfactions, embarrassment and complaints.

2 Validating explanations with Discourse Trees

2.1 Classes of explanation

To systematically treat the classes of explanation, we select an environment where customers receive explanations from customer service regarding certain dissatisfactions these customers encountered. If these customers are not satisfied with explanations, they frequently submit detailed complaints to consumer advocacy sites. In some of these complaints these customers explain why they are right and why the company's explanation is wrong. From these training sets we select the *good/bad* explanation pairs and define respective explanation classes via learning to recognize them.

Another way to consider a bad explanation is what we call an *explanation attempt*: a logical chain is built but it has some omissions and inconsistencies so that the explanation is bad. An absense of a logical chain means the absense of explanation; otherwise, if such chain obeys certain logical properties it can be interpreted by something else besides explanation but instead argumentation, clarification, confirmation or other mental or epistemic state.

2.2 Explanation and Argumentation

Explanations are correlated with argumentation and sentiments. A request to explain is usually associated with certain arguments and a negative sentiment.
For an arbitrary statement S a person may have little or no prior reason for believing this statement to be true. In this case a cognitive response is a doubt, which is articulated with a request for evidence.
Evidence is a kind of reason, and the attempt to provide evidence in support of a conclusion is normally called an *argument*. Argument reasoning is represented on the top of Fig. 1.

On the other hand a person may already know S and require no further evidence for the truth of S. But she still may not understand why S holds (occurred, happened etc. In this case she would request for a cause. Explanation is defined as an attempt to provide a cause in support of a conclusion. Explanation reasoning may be represented in the bottom of Fig. 1.

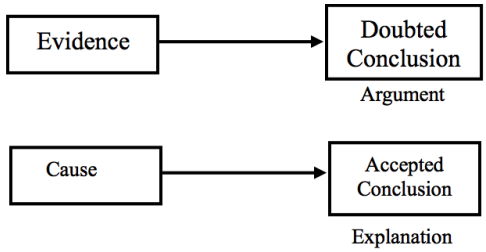

Fig.1: Relationship between argumentation and explanation

2.3 Hybrid discourse trees

In the banking domain *nonsufficient fund fee* (NSF) is a major problem that banks have difficulties communicating with customers. An example of brief, informal explanation follows:

It's not always easy to understand overdraft fees. When a transaction drops your checking account balance below zero, what happens next is up to your bank. A bank or credit union might pay for the transaction or decline it and, either way, could charge you a fee.

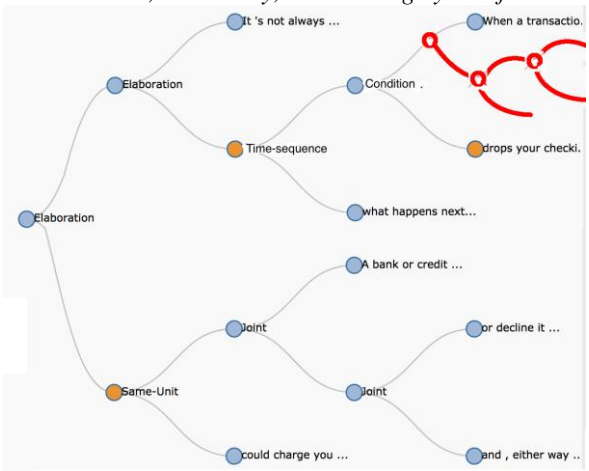

Fig 2. Discourse tree of explanation text with the imaginary part shown in the top-right for nucleus '*transaction*'.

The concept of *transaction* is not tackled in this text explaining nonsufficient fee. An ontology could specify that transaction = *{wiring, purchasing, sending money}* but it is hard to be complete. Instead, one can complement the notion of transaction via additional text that will elaborate on transaction, providing more details on it.

Hence *Elaboration* relation for nucleus *transaction* is not in actual DT but is assumed by a recipient of this explanation text. We refer to such rhetorical relations as Imaginary: they are not produced from text but are instead induced by the context of explanation. Such multiple imaginary RRs form additional nodes of an actual DT for a text being communicated. We refer to the extended DT as *complete*: it combines the actual DT and its imaginary parts. Naturally, the latter can be dependent on the recipient: different people keep in mind distinct instances of *transactions*.

We formalize this intuition by using discourse structure of the text expressed by DTs. Arcs of this tree correspond to rhetorical relations (RR), connecting text blocks called Elementary Discourse Units (EDU). We rely on the Rhetorical Structure Theory (RST, Mann and Thompson, 1988) when construct and describe discourse structure of the text.

When people explain stuff, they do not have to enumerate all premises: some of them implicitly occurring in the explanation chain and are assumed by the person providing explanation to be known or believed by an addressee. However, a DT for a text containing explanation only incudes EDUs from actual text and assumed, implicit parts with its entities and phrases (which are supposed to enter explanation sequence) are absent. How can we cover these implicit entities and phrases?

In the considered example *Elaboration* relation for nucleus *transaction* is not in actual CDT but is assumed by a recipient of this explanation text. We refer to such rhetorical relations as *Imaginary*: they are not produced from text but are instead induced by the context of explanation. Such multiple imaginary RRs form additional nodes of an actual DT for a text being communicated. We refer to the combined CDTs as *hybrid*: it combines the actual CDT and its imaginary parts. Naturally, the latter can be dependent on the recipient: different people keep in mind distinct instances of *transactions*. Complete discourse tree for the example is shown on Fig.2. Complete discourse trees also have communicative actions attached to their edges in the form of VerbNet verb signatures (Galitsky and Parnis, 2019).

2.4 Semantic representation

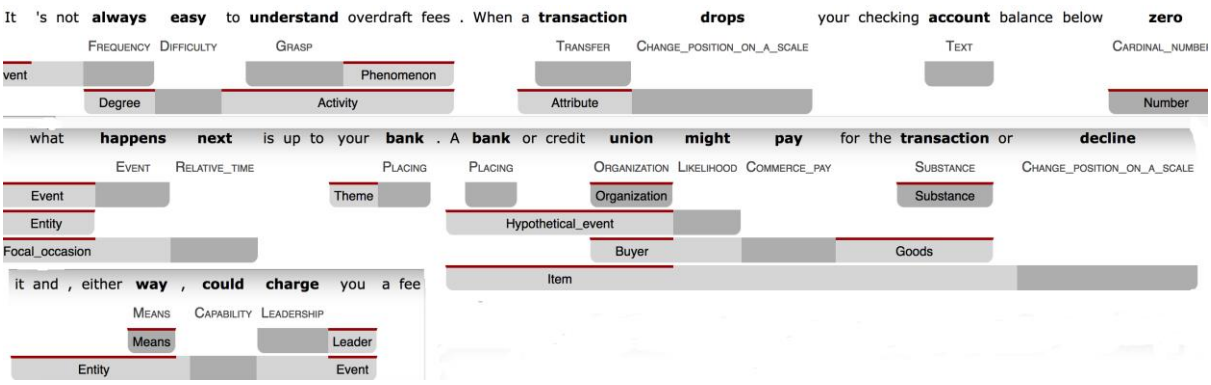

Fig.3 Frame semantic parse for the explanation

A frame semantic parse for the same text is shown in Fig. 3. The reader observes that it is hard to tag entities and determine context properly. *Bank* is tagged as Placing (not disambiguated properly) and '*credit union might*' is determined as a hypothetical event since *union* is represented literally, as an *organization*, separately from *credit*. Overall, the main expression being explained, '*transaction drops your checking account balance below zero',* is not represented as a cause of a problem by semantic analysis, since a higher level considerations involving a banking – related ontology would be required.

Instead of relying on semantic–level analysis to classify explanations, we propose a discourse-level machinery. This machinery allows including the explanation structure beyond the ones from explanation text but also from the accompanying texts mined from various sources to obtain a complete logical structure of the entities involved in explanation.

2.5 Discourse tree of explanations

Valid explanation in text follow certain rhetoric patterns. In addition to default relations of Elaborations, valid explanation relies on *Cause, Condition*, and domain-specific *Comparison* (Fig. 4) As an example, we provide an explanation for why *thunder sound comes after lightning*:

'We see the lightning before we hear the thunder. This is because light travels faster than sound. The light from the lightning comes to our eyes much quicker than the sound from the lightning. So we hear it later than we see it.'

```
joint
  elaboration (LeftToRight)
    cause (LeftToRight)
      temporal (LeftToRight)
        TEXT: We see the lightning
        TEXT: before we hear the thunder .
      TEXT: This is because light travels faster than sound .
    elaboration (LeftToRight)
      TEXT: The light from the lightning travels to our eyes much quicker than the sound from the
lightning .
       comparison (LeftToRight)
        TEXT:so we hear it later
        TEXT:than we see it .
```

Fig. 4: A discourse tree for an explanation of a lightning

The clause we need to obtain for an implication in the explanation chain is verb-group-for-moving {*moves, travels, comes*} *faster* → *verb-group-for-moving-result* {*earlier*}. This clause can be easily obtained by web mining, searching for expression 'if noun verb-group-for-moving *faster* then noun *verb-group-for-moving-result earlier.*

What would make this DT look like a one for invalid explanation? If any RR under top-level *Elaboration* turns into *Joint* it would mean that the explanation chain is interrupted.

We explore argumentation structure example of (Toulmin, 1958, Kennedy et al., 2006). We show two visualizations of the discourse tree and the explanation chain (in the middle) in Fig. 5.

```
  elaboration
    TEXT:Harry was born in Bermuda .
    explanation (LeftToRight)
      attribution (LeftToRight)
        TEXT:A person born in Bermuda is a British subject .
        TEXT:It is on account of the following statutes 123 .
      condition
        TEXT:So , presumably , Harry is a British subject ,
        Joint
          TEXT:unless both his parents were aliens ,

          TEXT:or he has become a naturalized American .
```

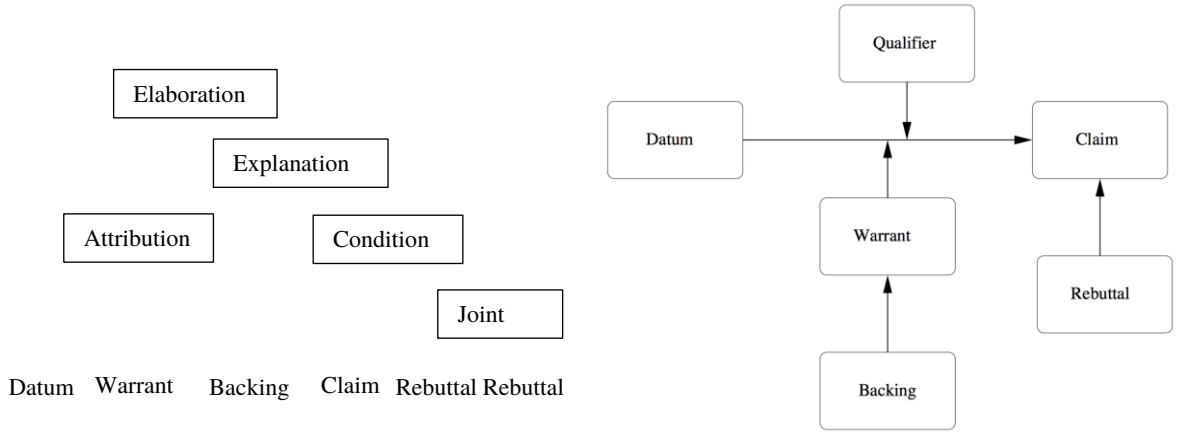

Fig. 5: Toulmin's argument structure (in the middle) and its rhetorical representation via EDUs (on the top) and via discourse relations (on the bottom)

An interesting application of Toulmin's model is the argumentative grammar by Lo Cascio (1991), a work that, by defining associative rules for argumentative acts, is naturally applicable, and indeed has been applied, to the analysis of discourse structure in the pre-DT times.

2.6 Logical Validation of Explanation via Discourse trees

Logically, explanation of text *S* is a chain of premises $P_1, \ldots, P_m$ which imply S. S is frequently referred to as a subject of explanation. For this chain $P_1, \ldots, P_m$ each element P_i is implied by its predecessors: $P_1, \ldots P_{i-1} \Rightarrow P_i$. In terms of a discourse tree, there should be a path in it where these implications are realized via rhetorical relations. We intend to define a mapping between EDUs of a DT and entities P_i occurring in these EDUs which form the explanation chain. In terms on underlying text, P_i are entities or phrases which can be represented as logical atoms or terms.

These implication-focused rhetorical relations *rr* are:

1) *elaboration*: P_i can be an elaboration of P_{i-1} ;
2) *attribution*: P_i can be attributed to P_{i-1} ;
3) *cause*: this is a most straightforward case,

Hence $P_i \Rightarrow P_j$ if $rr(EDU_i, EDU_j)$ where $P_i \in EDU_i$ and $P_j \in EDU_j$. We refer to this condition as "*explainability" via Discourse Tree*.

Actual sequence $P_1, \ldots, P_m$ for S is not known, but for each S we have a set of good explanations $P_{g1}, \ldots, P_{gm}$ and a set of bad explanations $P_{b1}, \ldots, P_{b2}$.

Good explanation sequences obey *explainability via DT* condition and bad – do not (Galitsky 2018). Bad explanation sequences might obey *explainability via DT* condition for some P_{bi}. If a DT for a text is such that *explainability via DT* condition does not hold for any P_{bi} then this DT does not include any explanation at all.

The reader can observe that to define a good and a bad explanation via a DT one needs a training set covering all involved entities and phrasing Pi occurring in both positive and negative training sets.

2.7 Constructing Imaginary Part of a Discourse Tree

By our definition imaginary DTs are the ones not obtained from actual text but instead built on demand to augment the actual ones. For a given chain $P_1, \ldots, P_i', \ldots, P_m$ let P_i' be the entity which is not explicitly mention in a text but instead is assumed to be known to the addressee. This P_i' should occur in other texts in a training dataset. To make the *explainability via DT* condition applicable, we need to augment actual DT_{actual} with imaginary $DT_{imaginary}$ such that $P_i' \in$ EDU of this $DT_{imaginary}$. We denote DT_{actual} $DT_{imaginary}$ as $DT_{complete}$.

If we have two textual explanations in the positive set of good explanations for the same S, T_1 and T_2:

T_1: P_1 ,..., $P_m \Rightarrow$ S
T_2: P_1 , P_i', ..., $P_m \Rightarrow$ S

then we can assume that P_i' should occur in a complete explanation for S and since it does not occur in T_1 then DT(T_1) should be augmented with $DT_{imaginary}$ such that $P_i' \in$ EDU of this $DT_{imaginary}$.

3 Learning Framework and Evaluation

In this section we automate our validation of text convincingness including description of a training dataset and learning framework.

We conduct our evaluation in two steps. Firstly, we try to distinguish between texts with explanation and without explanation. This task can be accomplished without an involvement of virtual DTs. Secondly, once we confirm that that can be done reasonably well, we drill into more specific tasks of differentiating between good and bad explanation chains within the dataset of the first task.

3.1 Building a Dataset of Good/bad Explanation Chains

We form the positive explanation dataset from the following sources:

1. Customer complaints;
2. Paragraphs from physics and biology textbook;
3. Yahoo! Answers for *Why/How-to* questions.

The negative training dataset includes the sources of a totally different nature:

1. Definition/factoid paragraphs from Wikipedia, usually, first paragraphs;
2. First paragraphs of news articles introducing new events;
3. Political news from Functional Text Dimension dataset.

We formed the balances components of the positive and negative dataset for both tasks: each component includes 240 short texts 5-8 sentences (250-400 words).

We now comment on each source. The purpose of the customer complaint dataset is to collect texts where authors do their best to explain their points across by employing all means to show that they are right and their opponents are wrong. Complaints are emotionally charged texts providing explanation of problems they encountered with a financial service, how they tried to explain their viewpoint to a company and also a description of how these customers attempted to solve it (Galitsky et al., 2008, GitHub Customer Complaints dataset 2019).

Also, to select types of text with and without explanation, we adopt the genre system and the corpora from (Lee, 2001). The genre system is constructed relying on the Functional Text Dimensions. These are genre annotations which reflect judgments as to what extent a text can be interpreted as belonging to a generalized functional category. A genre is a combination of several dimensions. For the positive dataset, we select the genre with the highest density of explanation such as scientific textbook. For the negative dataset, we focus on the genres which are least likely to contain explanations, such as advertisement, fiction-prose, instruction manuals and political news. The last one is chosen since it has the least likelihood to contain an explanation.

For the positive dataset for the second task, as good explanation chains, we rely on the following sources:

1. Customer complaints with *valid* argumentation patterns;
2. Paragraphs from phisics textbook explaining certain phenomena, which are neither factoid nor definitional;
3. Yahoo! Answers for *Why/How-to* questions;

We form the negative dataset from the following sources:

1. Customer complaints with *invalid* argumentation patterns; these complaints are inconsistent, illogical and rely on emotions to bring their points across;
2. Paragraphs from phisics textbook formulating longer questions and problems;
3. Yahoo! Answers for *Why* (not *How-to*) questions which are reduced to break the explanation flow. Sentences are deleted or re-shuffled to produce an incohesive, non-systematic explanation.

3.2 Crawling Information for Imaginary Discourse Tree Construction

Imaginary DTs can be found by employing background knowledge in a domain independent manner: no offline ontology construction is required. Documents that were found on the web can be the basis of constructing imaginary DTs following the algorithm described in the Section 2.4.

Given an actual part of the text A, we outline a top-level search strategy for finding a source for imaginary DTs (background knowledge) B.

1) Build DT for A;
2) Obtain pairs of entities from A that are not linked in DT (e.g. *thunder, eye*);
3) Obtain a set of search queries based on provided pairs of entities
4) For each query:
   a) Find a short list of candidate text fragments on the web using search engine API (such as Bing);
   b) Build DT for the text fragments;
   c) Select fragments which contain rhetoric relation (*Elaboration, Attribution, Cause*) linking this pair of entities;
   d) Choose the fragment with the highest relevance score

The *entity* mentioned in the algorithm can be interpreted in a few possible ways. It can be named entity, head of a noun phrase or a keyword extracted from a dataset.

Relevance score can be based on the score provided by the search engine. Another option – computing score based on structural discourse and syntactic similarity (Galitsky, 2017).

3.3 Learning Approaches and Pipelines

Discourse Tree Construction. A number of RST parsers constructing discourse tree of the text are available at the moments. For instance, in our previous studies we used the tool provided by (Surdeanu et.al., 2015) and (Joty et al., 2014).

Nearest Neighbor learning. To predict the label of the text, once the complete DT is built, one needs to compute its similarity with DTs for the positive class and verify that it is lower than similarity to the set of DTs for its negative class. Similarity between CDT's is defined by means of maximal common sub-DTs. Formal definitions of labeled graphs and domination relation on them used for construction of this operation can be found, e.g., in (Ganter, 2001).

SVM Tree Kernel learning. A DT can be represented by a vector of integer counts of each sub-tree type (without taking into account its ancestors). For Elementary Discourse Units (EDUs) as labels for terminal nodes only the phrase structure is retained: we suppose to label the terminal nodes with the sequence of phrase types instead of parse tree fragments. For the evaluation purpose Tree Kernel builder tool (Moschitti, 2006) can be used.

3.4 Detecting explanations and valid explanation chains

We first focus on the first task, detecting paragraphs of text which contain explanation, and estimate the detection rate in Table 1. We apply two different learning techniques, nearest neighbor (in the middle, greyed) and SVM TK, applied to the same discourse-level and syntactic data.

Table 1: Explanation detection rate

Source	P_{KNN}	R_{KNN}	$F1_{KNN}$	P_{SVM}	R_{SVM}	$F1_{SVM}$
1^+ vs 1^-	77.3	80.8	79.0	80.9	82.0	81.4
2^+ vs 2^-	78.6	76.4	77.5	74.6	74.8	74.7
3^+ vs 3^-	75.0	77.6	76.3	76.6	77.1	76.8
$1..3^+$ vs $1..3^-$	76.8	78.9	77.8	74.9	75.4	75.1

The highest recognition accuracy, reaching 80%, is achieved for the first pair of the dataset components, complaints vs wikipedia factois, most distinct 'intense' explanation vs enumeration of facts, with least explanations. The other datasets deliver 2-3% drop in recognition performance. These accuracies are comparable with various tasks in genre classification (one-against-all setting in Galitsky et al., 2016).

Table 2 shows the results of differentiation between good and bad explanation. The accuracy is about 12% lower than for the first task, since the difference between the good and bad explanation in text is fairly subtle.

Table 2: Recognizing good and bad explanation chains

Source	$P_{\text{-virtual}}$	$R_{\text{-virtual}}$	$F1_{\text{-virtual}}$	P	R	F1
1^+ vs 1^-	64.3	60.8	62.5	72.9	74.0	73.4
2^+ vs 2^-	68.2	65.9	67.0	74.6	74.8	74.7
3^+ vs 3^-	63.7	67.4	65.5	76.6	77.1	76.8
$1..3^+$ vs $1..3^-$	66.4	64.6	65.5	74.9	75.4	75.1

However, validation of explanation chain is an important task in a decision support. A low accuracy can still be leveraged by processing a large number of documents and detecting a birst in problematic explanation in a corpus of texts.

4 Discussion and Conclusions

In this work we considered a new approach to validating the convincingness of textual explanations. We introduced the notion of a *complete discourse tree* (complete DT) including actual and imaginary parts. Imaginary DT is constructed for the text about entities used but not explicitly defined in the actual text.

We outlined an algorithm for building an imaginary discourse tree. We also described a possible strategy for crawling background knowledge which is the source of the imaginary part. We also introduced the new dataset of good and bad explanations made by complainants in the financial domain. Finally, we outlined the learning framework used for automated detection of good and bad explanations. It is based on RST parsing and learning on complete discourse trees provided by the parser.

Both professional and non-professional writers provide explanations in texts but detection of invalid explanations is significantly harder in the former case compared to the latter. Professional writers in such domains as politics and business are capable of explaining “anything”, and in user-generated content errors are visible.

Detecting faulty explanations in user-generated content is important in automated Customer Relation Management systems where a response to user requests with valid explanation should be different to user response with invalid explanation.

It is important to combine rule-based learning frameworks with the ones with implicit feature engineering such as statistical and deep learning. The latest history of applications of statistical technique sheds a light on the limitation of these techniques for systematic exploration of a given domain. Once statistical learning delivered satisfactory results for discourse parsing, the interest to automated discourse analysis faded away. Since the researches in statistical ML for discourse parsing were mainly interested in recognition accuracies and not the interpretability of obtained DTs, no further attempts at leveraging obtained DTs were made. However, a number of studies including the given one demonstrate that DTs provide insights in the domain where keyword statistics does not help.

On the basis of work by Austin, Searle, Grice and Lorenzen, such discipline as pragmadialectics provides a comprehensive analysis of argumentative dialogues. This discipline combines the study on the formalism to represent data, from modern logic, and empirical observations, from descriptive linguistics, for the analysis of argumentative dialogues, modeled by dialectics, seen as sets of linguistic speech. The model proposes rule-base argumentative dialogues, but does not help with a dialogue generation algorithm.

Acknowledgements

The work of Dmitry Ilvovsky was supported by the Russian Science Foundation under grant 17-11-01294'.

References

Mann, William and Sandra Thompson. 1988. *Rhetorical structure theory: Towards a functional theory of text organization. Text* - Interdisciplinary Journal for the Study of Discourse, 8(3):243–281.

Explanation on Wikipedia. 2018. https://en.wikipedia.org/wiki/Explanation#Meta-explanation.

Galitsky, B., Kuznetsov SO. 2008. Learning communicative actions of conflicting human agents. J. Exp. Theor. Artif. Intell. 20(4): 277-317.

Galitsky B (2018) Customers' Retention Requires an Explainability Feature in Machine Learning Systems They Use. AAAI Spring Symposium Series.

Jansen, P., M. Surdeanu, and P. Clark. 2014. Discourse Complements Lexical Semantics for Nonfactoid Answer Reranking. ACL.

Surdeanu, M., Thomas Hicks, and Marco A. Valenzuela-Escarcega. 2015. Two Practical Rhetorical Structure Theory Parsers. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: Software Demonstrations (NAACL HLT), 2015.

Galitsky, B, D Ilvovsky, SO Kuznetsov (2016) Style and Genre Classification by Means of Deep Textual Parsing. Computational Linguistics and Intellectual Technologies: DIALOG, Moscow, Russia.

Galitsky, B. 2017. Matching parse thickets for open domain question answering, Data & Knowledge Engineering, Volume 107, January 2017, Pages 24-50.

Galitsky B, D Ilvovsky, SO Kuznetsov (2018) Detecting logical argumentation in text via communicative discourse tree. Journal of Experimental & Theoretical Artificial Intelligence 30 (5), 637-663.

Galitsky B, Parnis A (2018) Accessing Validity of Argumentation of Agents of the Internet of Everything. Artificial Intelligence for the Internet of Everything, 187-216.

Joty, S., Moschitti, A. 2014 Discriminative reranking of discourse parses using tree kernels. EMNLP 2014.

Ganter, B., Kuznetsov, S.O. 2001. Pattern structures and their projections. In: International Conference on Conceptual Structures. pp. 129-142. Springer.

Moschitti, A. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany.

Paul E. Dunne and Trevor J. M. Bench-Capon. 2006. Computational Models of Argument: Proceedings of COMMA 2006, IOS Press, 2006.

Lo Cascio, V. 1991. Grammatica dell'Argomentare: strategie e strutture [A grammar of Arguing: strategies and structures]. Firenze: La Nuova Italia.

Walton, D. 2007. Dialogical Models of Explanation. Explanation-Aware Computing: Papers from the 2007 AAAI Workshop, Association for the Advancement of Artificial Intelligence, Technical Report WS-07-06, AAAI Press, 2007,1-9.

Walton, D., Reed, C., Macagno, F. 2008. Argumentation schemes. Cambridge University Press

Lee, David YW. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. (2001)

Kennedy, X.J., Dorothy M. Kennedy, and Jane E. Aaron.. "Reasoning". The Bedford Reader. 9th ed. New York: Bedford/St. Martin's, 2006. p. 519–522.

Toulmin, S. The Uses of Argument. Cambridge At the University Press, 1958.