The Metalogue Debate Trainee Corpus: Data Collection and Annotations

Volha Petukhova¹, Andrei Malchanau¹, Youssef Oualil¹, Dietrich Klakow¹, Saturnino Luz², Fasih Haider³, Nick Campbell³, Dimitris Koryzis⁵, Dimitris Spiliotopoulos⁵, Pierre Albert², Nicklas Linz⁶, Jan Alexandersson⁶

¹Saarland University, Spoken Language Systems, Saarbrücken, Germany
²The University of Edinburgh, UK; ³Trinity College Dublin, Ireland
⁴Hellenic Parliament, Greece; ⁵University of Houston, TX, USA
⁶German Research Center for Artificial Intelligence (DFKI), Germany

Abstract

This paper describes the Metalogue Debate Trainee Corpus (DTC). DTC has been collected and annotated in order to facilitate the design of instructional and interactive models for Virtual Debate Coach application - an intelligent tutoring system used by young parliamentarians to train their debate skills. The training is concerned with the use of appropriate multimodal rhetorical devices in order to improve (1) the organization of arguments, (2) arguments’ content selection, and (3) argument delivery techniques. DTC contains tracking data from motion and speech capturing devices and semantic annotations - dialogue acts - as defined in ISO 24617-2 and discourse relations as defined in ISO 24617-8. The corpus comes with a manual describing the data collection process, annotation activities including an overview of basic concepts and their definitions including annotation schemes and guidelines on how to apply them, tools and other resources. DTC will be released in the ELRA catalogue in second half of 2018.

Keywords: debate argumentation, multimodal data collection, ISO standard annotations

1. Introduction

We currently observe a steadily growing interest of researchers and practitioners in natural argumentation modelling and in developing argumentation technologies. There are systems developed and deployed for legal domains to assist the lawyer in his search for similar past cases. (Teufel, 1999; Brüninghaus and Ashley, 2005), for mining arguments in social media with the goal to predict consumers sentiment (Bat, 2011), to analyse opinions in public discussions (Murakami and Raymond, 2010), to study citizen engagement (Purpura et al., 2008) and to recognize stance in political online debates (Somasundaran and Wiebe, 2010). Argumentation constitutes an important component of human intelligence. Educational studies have shown that constructing arguments and engaging in argumentative discussion enhance conceptual understanding of the subject matter (Wiley and Voss, 1999; Zohar and Nemet, 2002). Argumentation training systems are designed for the legal domains, e.g. to training hypothetical reasoning (Ashley et al., 2007). The TruthMapping® web application facilitates collaborative learning through argumentation. DebateGraph® used to train how to prevent opinion manipulation making inconsistent arguments. These and other developments were supported by corpora collection (Lawrence et al., 2012b) of political debates on internet forums, consisting of about 11,000 discussions and 390,000 posts. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness among others. The Yahoo News Annotated Comments Corpus (Napoles et al., 2017) is one of the largest annotated corpora of online human argumentative dialogues, with the most detailed set of annotations to identify argumentative, respectful exchanges containing persuasive, informative, and/or sympathetic comments. Larger projects have been used successfully as resources to study written and spoken argumentative discourse, e.g. Online Debate Forum® CE-EMNLP-2015, also known as IBM corpus, a selection of annotated arguments from Wikipedia articles (Rinott et al., 2015), documents of the European Court of Human Rights® UK Youth Parliament (UKYP) debates (Petukhova et al., 2016), the American Presidency Project (APP® and many more. For the application designed within the Metalogue project - Virtual Debate Coach - an interactive system used to train young parliamentarians to debate efficiently (Petukhova et al., 2017b), the Debate Trainees Corpus (DTC) of ‘natural’ multimodal arguments was collected. Trainees were trained to make choices from a wide range of rhetorical, lexical, syntactic, pragmatic and prosodic devices to deliver strong persuasive speeches (Petukhova et al., 2017c).

¹{v.petukhova,andrei.malchanau,youssef.oualil,dietrich.klakow}@lsv.uni-saarland.de
²{s.luz,pierre.albert}@ed.ac.uk;³{haiderf,nick}@tcd.ie;⁴dkoryzis@parliament.gr
⁵dspiliotopoulos@uh.edu;⁶{nicklas.linz,jan.alexandersson}@dfki.de

References

Ashley, A., et al. (2007). The TruthMapping® web application facilitates collaborative learning through argumentation. DebateGraph® used to train how to prevent opinion manipulation making inconsistent arguments. These and other developments were supported by corpora collection (Lawrence et al., 2012b) of political debates on internet forums, consisting of about 11,000 discussions and 390,000 posts. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness among others. The Yahoo News Annotated Comments Corpus (Napoles et al., 2017) is one of the largest annotated corpora of online human argumentative dialogues, with the most detailed set of annotations to identify argumentative, respectful exchanges containing persuasive, informative, and/or sympathetic comments. Larger projects have been used successfully as resources to study written and spoken argumentative discourse, e.g. Online Debate Forum® CE-EMNLP-2015, also known as IBM corpus, a selection of annotated arguments from Wikipedia articles (Rinott et al., 2015), documents of the European Court of Human Rights® UK Youth Parliament (UKYP) debates (Petukhova et al., 2016), the American Presidency Project (APP® and many more. For the application designed within the Metalogue project - Virtual Debate Coach - an interactive system used to train young parliamentarians to debate efficiently (Petukhova et al., 2017b), the Debate Trainees Corpus (DTC) of ‘natural’ multimodal arguments was collected. Trainees were trained to make choices from a wide range of rhetorical, lexical, syntactic, pragmatic and prosodic devices to deliver strong persuasive speeches (Petukhova et al., 2017c).
This paper describes the semantically annotated debate data collection undertaken within the project. The corpus will be delivered with audio and video recordings, Kinect tracking data, automatic and manual transcriptions, ISO compliant semantic annotations. Annotations will be provided in Anvil specific XML format, and converted to ISO 24617-2 Dialogue Act MarkUp Language (DiAML) (Bunt et al., 2012).

The corpus guidelines, reports, annotation schemes as well as data collection instructional material will be also provided to enable the replication of the carried out experiments.

2. Training Debate Argumentation

Debates, in particular political debates, constitute a large portion of public speeches. Skilled professional debaters give the impression that they truly believe what they say, know how to catch and keep the attention of the audience, and express authority, confidence, respect and friendliness. People generally associate certain speech, personality and interaction features with what they think is a ‘good public speaker’, see e.g. (Strangert and Deschamps, 2006).

The training of debate skills typically involves ad-hoc face-to-face classroom debates. The debater’s skills proficiency level is often judged on three criteria: (1) argument organization, (2) argument content, and (3) argument delivery. A debate is a communication process in which participants argue for or against a certain position proposed for the dispute. In a parliament setting, such initial position is called motion. In our training scenario, each debate session is motivated by a motion - new law proposal or changes to an existing law. A session consists of one or multiple training rounds, e.g. our session comprised four debate rounds, featuring different goals assigned to trainees by a Moderator (or Admin). Moderators initiate and further formally regulate the session(s). One or more Tutors attend the session and provide feedback to Trainees. Tutoring interventions are expected to inform trainees of mistakes, propose corrections, provide instructions, initiate ‘try again’ rounds, or highlight trainees’ successes. This involves immediate real-time ‘in-action’ and summative ‘about-action’ feedback (Schön, 1983) on the three debate aspects mentioned.
### Table 1: Example of participants’ minimal goals in own debate round.

<table>
<thead>
<tr>
<th>Issue under debate</th>
<th>Trainees minimum goals to achieve</th>
<th>Proponent</th>
<th>Opponent (conservative)</th>
<th>Opponent (liberal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking ban scope</td>
<td>All public places should be affected, allow smoking in bars and restaurants and open air areas like outside buildings, parks and beaches</td>
<td>Not all public places should be affected, allow smoking in bars and restaurants and open air areas like outside buildings, parks and beaches</td>
<td>Forbid smoking in all public spaces, special smoking areas outside buildings</td>
<td>Allow smoking in special areas in bars and restaurants, open air places also need smoking areas</td>
</tr>
<tr>
<td>Tobacco prices</td>
<td>Tobacco prices already high, increase no more than 3% a year</td>
<td>Tobacco prices are low, increase by 10% a year</td>
<td>Tobacco prices are still too attractive, increase by 5% a year</td>
<td>Tobacco prices are still too attractive, increase by 5% a year</td>
</tr>
<tr>
<td>Access to tobacco</td>
<td>Tobacco sold in supermarkets, specialized licensed tobacco shops, in bars and restaurants, and vending machines on street with secured buyer’s age control</td>
<td>Tobacco should be sold only in special licensed tobacco shops</td>
<td>Tobacco sold in supermarkets but hidden in special containers, prohibited to sell around schools (5km distance) and not available in bar or street vending machines</td>
<td></td>
</tr>
<tr>
<td>State control</td>
<td>No police control but municipal and administrative control, no penalties but warnings for the 1st time, repeated disobedience may be punished with penalties</td>
<td>Strong police presence in public places and penalties without warnings</td>
<td>No police control, municipal and administrative control, 1st time disobedience gets warning; second time penalties</td>
<td></td>
</tr>
<tr>
<td>Anti-smoking campaign</td>
<td>on TV (state channels 20 min broadcasting time a week); posters in every public place; ‘educated’ slogans on cigarettes; big newspapers 5 lines a week on the first 2-3 pages</td>
<td>on TV (all channels 30 min broadcasting time a week + one documentary a month); posters in every public place; slogans and scaring images on cigarettes; big newspapers 10 lines a week on the bottom of the front page</td>
<td>on TV (state channels 20 min broadcasting time a week); posters in every public place; ‘educated’ slogans on cigarettes; big newspapers 10 lines a week on the first 2-3 pages</td>
<td></td>
</tr>
</tbody>
</table>

above. The actual debate training session starts by the Proponent presenting the motion and an argument in favor of it. An argument is defined as consisting of a statement that can be supported by evidence. A statement (claim) is an assertion that deserves attention. There may be a conclusion which presents a result, which can be derived from certain evidence (premises). An argument has certain internal structure and correspond to a discourse unit, often called Argumentative Discourse Unit (ADU, see e.g. (Peldszus and Stede, 2013). Claim and premises are dialogue acts related to each other by means of discourse relations, see Section 5.2. The task of the Opponent is to attack the proponent’s argument by rebutting parts of the argument (premises) or its conclusion, or by undercutting their supporting inference(-s). Both trainees can be in the role of either an proponent or opponent. Moderator(-s) may terminate debating and collect trainees feedback about the debate process and their progress in achieving goals assigned to them. Figure[1] presents the conceptual class diagram for a Metagleue debate training session.

### 3. Scenario and Data Collection

The specific setting considered for the data collection involves a debate scenario about anti-smoking legislation in Greece. The initial proposal for a smoking ban is supported by the proposing (governmental) party. The goal of the proposer is to get a majority vote while agreeing on as few amendments as possible.

Our core data collection activity involved debate trainees, school children aged 14-15 years who have been exposed to little debate training. A session involved a pair of participants: one assigned the role of proposer, the other the role of either liberal or conservative opponent. Each participant was given a set of minimal goals concerning: (1) the total ban on smoking in public spaces; (1) limiting youth access to tobacco products; (3) improving the effectiveness of anti-smoking campaign; (4) state control and reinforcement policy; (5) and raising prices on tobacco products. Participants were not allowed to disclose their goals to the other parties prior to the interaction. Three human tutors evaluated debate performance. Table[1] provides an example of minimal goals that trainees playing different roles should achieve in one debate round.

The collected data consists of 12 sessions with a duration of 2.5 hours, comprising 400 arguments (Argumentative Discourse Units, ADUs) from 6 different bilingual English/Greek speakers.

### 4. Multimodal Recording and Synchronisation

Training sessions were recorded in a quiet room under special lighting conditions, ensuring that there were no windows behind the participants and that the participants’ faces were not in shadow. Two Kinect V1 sensors, each facing one participant as much as possible, were placed at a distance of 1.5-2m to the participants. A Kinect V2 sensor was also used to track both participants. Body and face tracking data were stored in an XML format containing elements for frames, faces, joint orientation and bone rotation with respect to the camera’s coordinates. Participants faced each other, and markers were placed on the floor to constrain the participants to a limited area. In addition to Kinect’s videos, the recordings included two separate video streams, recorded by conventional video cameras. Figure[2] depicts the technical setup for Metagleue debate sessions.

Speech was captured by two audio Tascam Dr-40 recorders and saved in MS WAV format. Speech files are of two types: (1) full dialogue session recorded per speaker, and (2) cut audio files per speaker and roughly per turn (after speaker diarization). Speaker diarization has been partly carried out manually using the Audacity tool and partly automatically using LIUM tool (Rouvier et al., 2013). The speech signal files contain timestamps - start and end time

---

[1]For more details on segmentation and annotation performed, we refer to [Petukhova et al., 2016].

[2]The recordings were performed in the following setting: sample rate (48KHz), sample size (16-bit), sample format (linear PCM) with stereo channel which was later converted to mono .

Figure 2: Recording set up for Metalogue debate sessions, adapted from Haider et al. (2017).

![Recording set up diagram]

Figure 3: Argument structure observed in Metalogue debate data.

- and additional comments on acoustic and temporal conditions (noise, long silences, etc.) in the file name. For example, 08.22-08.30.n.wav is the segment which started at 8 minutes and 22 seconds and finished at 8 minutes and 30 seconds during the recording session; and it contains some noise indicated by “n”.

The Kinect and video streams were synchronised with audio using the Final Cut Pro X software. The resulting media were converted to a ANVIL compatible format. Participants’ speech has been transcribed semi-automatically by (1) running the Automatic Speech Recognizer (ASR) Kaldi (Povey, 2011) and (2) correcting ASR output manually. Corrected transcriptions are used to re-train/improve language models. All types of transcriptions are stored in plain text and converted to TEI compliant format (ISO, 2006).

5. Annotation Design

Organization of arguments is the planning and preparation involving Argument as a general claim, Reason(-s) and Evidence. This structure is often called ARE.

Good debaters are distinguished by concise clear arguments and try to make their arguments understandable for their addresseees. For this purpose, debaters often use linguistic cues such as discourse markers and meta-discoursive acts. For example, 'I will talk in favour of ... Because ... Since according to the international research shows...'. Thus, discourse relations between two or more dialogue acts (argument’s premises or conclusions) are often marked explicitly by means of discourse markers to support Justification, Motivation, Cause/Result, Background/Evaluation, Evidence and Circumstance links. Figure 3 depicts the most frequently observed of the Metalogue arguments, about 80% of data accounts for this pattern. The main claim, i.e. Statement, is supported by either a Reason or Evidence, and is wrapped up by a Re-Statement in the form of a Summary or Conclusion. For example:

(1) D1

5.1. Dialogue Acts

In Metalogue, we mostly consider annotations of semantic and pragmatic multimodal phenomena. For this purposes, dialogue acts play an important role. The ISO 24617-2 dialogue act annotation standard is used which allows the analysis of dialogue behaviour as having communicative func-

---

13 For more details on the data collection and synchronisation process we refer to (Haider et al., 2017).

14 It should be noticed that the corpus contains a significant proportion of non-native English speakers, varying in fluency from nearly-native to challenging-to-transcribe.

15 See http://www.slideshare.net/Cherye/advanced-debating-techniques and Petukhova et al.,

16 Crismore et al., 1993 define metadiscourse as “linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organize, interpret and evaluate the information given”, e.g. Shifting Topic, Marking Asides, etc.

17 Here and henceforth Dk stands for Debater k; the subscript is the index of the identified dialogue act.
5.2. Discourse Relations

Discourse relations were annotated using the annotation scheme designed for the Penn Discourse TreeBank (DPTB) corpus (Prasad et al., 2008), extended with discourse segment relations from the taxonomy proposed in (Hovy and Maier, 1995). Table 2 presents the types and frequencies of the relations along with the inter-annotator agreement reached annotating each relation type. For relations like Elaboration, Evidence, Justification, Reason, Conclude and Restate, which are important for the debate argument identification and processing, a substantial agreement has been achieved. The annotated discourse relations were mapped to those defined in ISO 24617-8 standard, which was published after all DTC sessions were annotated.

5.3. Argumentative Discourse Units

We segmented debates into Argumentative Discourse Units (ADUs), defined as a unit which consists of one or more premises and one conclusion, possibly restated or paraphrased several times by the same speaker. To identify ADUs, we followed the approach proposed by (Peldszus and Stede, 2013), who suggest to first segment into Elementary Discourse Units (EDUs) as minimal discourse building blocks, then establish relationships between two or more EDUs, and combine those into ADUs. Identifying ADUs, we observed a very frequent pattern: an ADU will mostly start with a simple Inform act and end when an Inform Conclude or Restatement is identified, or before another Inform act is performed by the same speaker which is not involved in any discourse relation, see Figure 3 and example in (1), or another speaker claimed the turn. Finally, to capture support and attack links between arguments produced by different speakers, we identified explicit and implicit agreement and disagreement dialogue acts signalling support or attack of arguments through the functional dependence relations defined in (ISO, 2012) between the detected argument conclusions. For example:

(2) D1_47;D1_1_2: The government should launch effective anti-smoking campaign before it’s too late [Inform] D2_4;D2_2_1: Decision to smoke or not is the personal issue and state shouldn’t interfere [Inform&Disagreement D1_47] - Attack D1_1_2 D7_2;D7_2_1: I think public health is one of the most important tasks that the government should perform [Inform&Agreement D1_47 & Disagreement D2_1] - Support D1_1_2/Attack D2_2_1

Debater 1 states that an anti-smoking campaign is needed and it is the government responsibility. Debater 2 thinks that smoking is the personal responsibility and government should not interfere. Debater 7 supports argument 1.2 and thereby attacks the arguments 2.1. These links are modelled as part of the debaters’ information states, see (Petukhova et al., 2016).

5.4. Dialogue Act Markup Language

ISO standard 24617-2 includes the definition of the Dialogue Act Markup Language (DiAML). The representation of a dialogue act annotation makes use of the XML element <dialogueAct> with attributes such as speaker, addressee, communicative function, dimension, qualifiers, dependence relations. Additionally, rhetorical (discourse) relations among dialogue acts are represented by means of <rhetolink> elements. All these types are defined in diaml namespace in the defined DIAML_Types.xsd scheme. In DiAML_Containers.xsd elements such as primary data tokens and sounds, and functional segments are specified without a namespace. The last allows to reuse DiAML_Containers in other (not-diaml) schema that may also specify domain-dependent semantics. Such semantics is usually a description of all possible elements of the semantic content of dialogue acts. Semantics can be specified to represent predicate-argument structures, named en-

Table 2: Distribution of Inform acts connected by a discourse relation in the corpus (* defined in DPTB; ** defined by Hovy and Maier, 1995; *** in both taxonomies).

<table>
<thead>
<tr>
<th>Discourse relation</th>
<th>Relative frequency (in %)</th>
<th>Cohen’s kappa scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaboration**</td>
<td>28.1</td>
<td>0.67</td>
</tr>
<tr>
<td>Evidence**</td>
<td>21.4</td>
<td>0.72</td>
</tr>
<tr>
<td>Justify***</td>
<td>16.1</td>
<td>0.76</td>
</tr>
<tr>
<td>Condition***</td>
<td>0.7</td>
<td>0.34</td>
</tr>
<tr>
<td>Motivation***</td>
<td>1.4</td>
<td>0.48</td>
</tr>
<tr>
<td>Background***</td>
<td>0.3</td>
<td>0.18</td>
</tr>
<tr>
<td>Cause***</td>
<td>3.4</td>
<td>0.37</td>
</tr>
<tr>
<td>Result***</td>
<td>2.2</td>
<td>0.26</td>
</tr>
<tr>
<td>Reason*</td>
<td>10.6</td>
<td>0.73</td>
</tr>
<tr>
<td>Conclude**</td>
<td>5.7</td>
<td>0.71</td>
</tr>
<tr>
<td>Restatement***</td>
<td>10.1</td>
<td>0.76</td>
</tr>
</tbody>
</table>

http://dit.uvt.nl/
### Table 3: Metalogue Debate Trainee Corpus overview.

<table>
<thead>
<tr>
<th>Type</th>
<th>Content</th>
<th>Format</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debate minimal goals cards</td>
<td>4 rounds</td>
<td>pdf</td>
<td>defined for Proponent and Opponent</td>
</tr>
<tr>
<td>Metadata</td>
<td>participants (id, native language sex, age at collection)</td>
<td>xml, TEI compliant</td>
<td>generated from participants forms</td>
</tr>
<tr>
<td>Signals</td>
<td>sound recordings</td>
<td>mono, 96000Hz sample rate</td>
<td>1 channel per speaker</td>
</tr>
<tr>
<td></td>
<td>wav files</td>
<td>24-bit sample format</td>
<td>cut per speaker/turn</td>
</tr>
<tr>
<td></td>
<td>Kinect tracking</td>
<td>xml, 30 frames per second</td>
<td>tracked per speaker/turn recorded per speaker</td>
</tr>
<tr>
<td>Automatic Speech Recognition Transcriptions</td>
<td>turn (id, start, end, string)</td>
<td>plain text</td>
<td>automatic</td>
</tr>
<tr>
<td></td>
<td>turn (id, start, end, string)</td>
<td>plain text</td>
<td>manual</td>
</tr>
<tr>
<td></td>
<td>utterance (id, start, end, string)</td>
<td>xml, TEI compliant</td>
<td>automatic</td>
</tr>
<tr>
<td></td>
<td>functional segments (id, start, end, pointers)</td>
<td>xml, TEI compliant</td>
<td>automatic</td>
</tr>
<tr>
<td>DA annotations</td>
<td>dialogue act (sender, dimension, communicative function, qualifier functionalDependenceRelation feedbackDependenceRelation rhetoricalLinks)</td>
<td>Anvil and DiAML</td>
<td>manual</td>
</tr>
</tbody>
</table>

6. Corpus Overview

The Metalogue DTC corpus comprises signals, tracking data, transcriptions, meta-data, semantic and pragmatic annotations in standard xml-format. Table 3 provides corpora overview specifying type of data planned for release.

Six types of semantic annotations were performed by two trained and one expert annotators. In total, the Metalogue DTC corpus contains about 10,000 annotated entities. The Metalogue corpus will be published in the ELRA catalogue and be available to the community for research purposes in 2018.

7. Acknowledgements

The reported research is partly funded by the EU FP7 Metalogue project, under grant agreement number: 611073.

8. Bibliographical References


Napoles, C., Tetreault, J., Pappu, A., Rosato, E., and

---

21 See also [Petukhova et al., 2017a](http://catalog.elra.info/).


