The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues

Volha Petukhova[1], Martin Gropp[1], Dietrich Klakow[1], Anna Schmidt[1],
Gregor Eigner[2], Mario Topf[2], Stefan Srb[2],
Petr Motlicek[3], Blaise Potard[3], John Dines[4],
Olivier Deroo[5], Ronny Egeler[6], Uwe Meinz[6], Steffen Liersch[6]

[1] Saarland University, Spoken Language Systems, Saarbrücken, Germany
[2] Mipumi Games GmbH, Vienna, Austria
[3] Idiap Research Institute, Martigny, Switzerland
[4] KOMEI SA, Switzerland
[5] Acapela Group S.A., Mons, Belgium
[6] Sikom Software GmbH, Heidelberg, Germany
[1]{v.petukhova,martin.gropp,dietrich.klakow,anna.schmidt}@lsv.uni-saarland.de;
[2]{g.eigner,m.topf,s.srb}@mipumi.com; [3]{motlicek,blaise.potard}@idiap.ch;
[4]john.dines@koemei.com; [5]olivier.deroo@acapela-group.com;
[6]{r.egeler,u.meinz, s.liersch}@sikom.de

Abstract

The paper describes a project for continuous data collection for a spoken dialogue system engaged in Question-Answering interactions in English. The Wizard-of-Oz method used in the bootstrap phase is presented, and several types of resulting dialogue annotations are described. The resulting corpus will be publicly released.

Keywords: continuous dialogue data collection, Wizard-of-Oz experiments, semantic annotations

1. Introduction

This paper describes the data collection and annotation activities carried out within the DBOX project[1]. This project aims to develop interactive games based on spoken natural language human-computer dialogues, in 3 European languages: English, German and French.

A common procedure in design of human-computer dialogue systems is, first, to collect human-human data in order to model natural human dialogue behaviour, for better understanding of phenomena of human interactions and predicting interlocutors actions, and then to develop dialogue system components.

There are numerous dialogue corpora collected for different domains and applications. For instance:

- *MapTask* dialogue corpus[2] a collection of instructing dialogues where one participant plays the role of an instruction-giver while another participant, the instruction-follower, navigates through a map.
- *AMI* dialogue corpus[3] a 100-hours meeting corpus collected in 2005, containing human-human multiparty interactions in English.
- *Switchboard* dialogue corpus[4] that consists of 650 spontaneous telephone conversations between speakers of American English.
- *Coconut* dialogue corpus[5] a collection of 35 two-party negotiation typed computer-mediated dialogues.

Nevertheless, for the first DBOX gaming scenario, "Quiz game", we failed to find any suitable dialogue data. Hence, the decision has been made by the consortium to collect spoken dialogues for this scenario. Building a dialogue corpus is a very taxing activity, especially when it requires human participants involvement in data collection and manual annotations. In oder to reduce these efforts, the *Wizard-of-Oz* methodology is often used (Dahlbäck et al., 1993). In such experiments, the dialogue system is usually replaced by a human *Wizard* who simulates the system's behaviour by acting according to a pre-defined script.

Another alternative is to use *simulated users*. With good user modeling, a dialogue system could be rapidly prototyped and evaluated. However, it is very challenging to ensure that the user model truly reflects what real users are likely to do, which is often dependent on very subtle aspects of the dialogue design and task domain (Paek, 2006). Thus, some initial real human data is required anyway in order to further simulate further user behaviour.

The third method, which has been employed to collect the DBOX data, is a *continuous data collection*, where we first

[1] DBOX is an Eureka project, number E! 7152 http://www.idiap.ch/project/d-box/

[2] For example, http://www.hcrc.ed.ac.uk/maptask/ for English; http://www1.uni-hamburg.de/exmaralda/files/z2-hamatac/public/index.html for German; http://crdo.up.univ-aix.fr/voir_depot.php?lang=en&id=732&prefix=sldr (MAPTASK-AIX) for French, and many other languages.

[3] http://www.amiproject.org/

[4] http://groups.inf.ed.ac.uk/switchboard/

[5] http://www.pitt.edu/~coconut/coconut-corpus.html

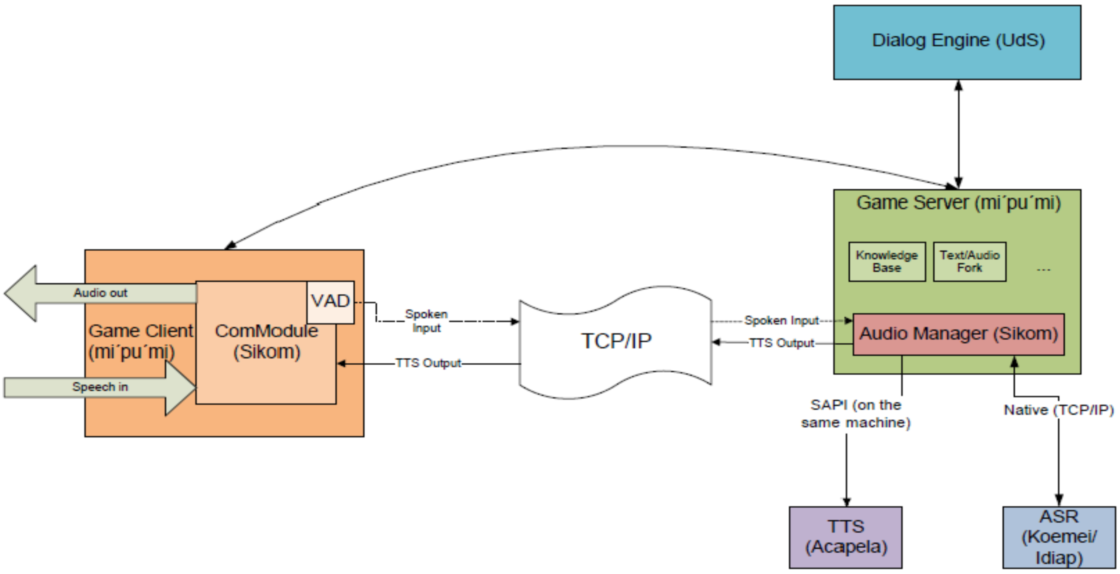

Figure 1: Overall DBOX system architecture.

start with Wizard-of-Oz scripted experiments, and then replace the human Wizard by an increasingly advanced dialogue system, using evaluation data for system improvement in each iteration.

The paper is structured as follows. Section 2 explains the DBOX scenario and provides an overview of the targeted dialogue/game phenomena that are included in the Wizard script. Section 3 describes the collection of dialogue and non-dialogue data. Section 4 discusses the performed annotations, with indication of reliability of the defined and existing annotation schemes in terms of inter-annotator agreement. Section 5 concludes the reported work by summarizing corpus collection and data annotation activities, and outlines future research.

2. Scenario and targeted dialogue phenomena

The first DBOX scenario is concerned with a Quiz Game. The game is comparable to the famous US TV show of the 70s-80s 'What's my line?' or the German equivalent 'Was bin ich?'. In these games a famous person is hidden from the players, whose task is to guess his/her name by asking 'Yes/No Questions' (no more than 10 in one round). Whoever is the first to guess the name correctly – wins. The main difference with our scenario is that our players are allowed to ask any type of questions, e.g., set Questions such as 'Where were you born?' or 'What are you famous for?'. The designed system, thus, provides an interactive game where the system holds the facts about a famous person's life, and the user's task is to guess his/her name by asking ten questions of various types. The system prevents the user to ask direct questions about the name or an alias.

The dialogue system relies on a Question-Answering (QA) approach in the sense that its main task is to understand users' questions, find and retrieve an answer from the stored description, and return it to the user. However, what is more important for our purposes is to build an *interactive Question-Answering Dialogue System* (QADS) where the answers are extracted not as information chunks or slot fillers, but rather form full-fledged dialogue utterances. Moreover, the system needs to show an interactive gaming behaviour that is natural to its users, e.g., provide feedback, manage turns, time and contact (dependent on the quality of a communication channel), produce some social signals and acts, e.g., encouraging vs. downplaying, polite vs. rude, positive vs. negative attitude towards players or their actions, etc. All these would make a game more entertaining. Thus, the following dialogue and gaming phenomena need to be elicited when collecting the data:

- dialogue/game opening and greetings, e.g. 'Nice to meet you', 'Thank you for playing with me'; or 'Welcome back!','How are you doing today?' for returning users;
- elaborative feedback, e.g. 'This is a nice question', 'Clever move!' or 'Not your day today';
- modal expression uttering (un)certainty 'That might be', 'Certainly not', 'I hope I can tell so', etc.;
- check questions 'Am I correct in assuming you asked whether...';
- requests to repeat 'Sorry, I didn't understand/hear you, could you repeat?';
- hesitations and stallings 'Wait a second', 'Let me think', 'Uh Um';
- closings 'You did very well!', 'Wonderful!', 'Congratulations!', 'You are a profi!', 'Good job though', 'Next time better!', and many more.

An example of a collected dialogue can be found in Appendix.

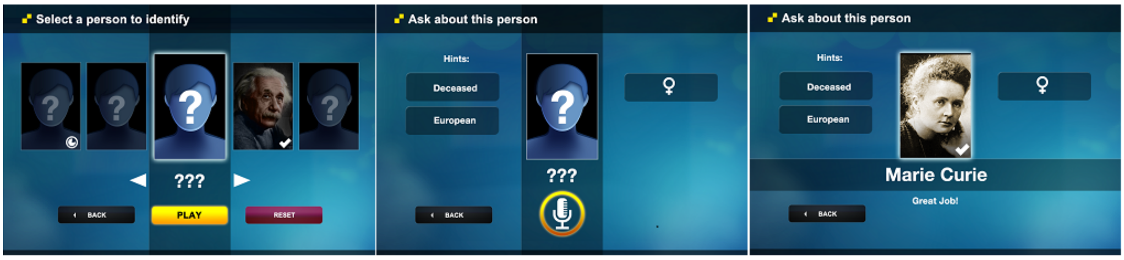

Figure 2: DBOX Quiz game interface

3. Continuous data collection

As we briefly mentioned in the introduction, the data is collected continuously during the whole project life. Preliminary human-human data has been collected in a Wizard-of-Oz setting in parallel with the spoken dialogue system design, which obviously at this early project stage is very simple (scripted). The system will be replaced in later stages by a more advanced one and evaluated. Data collected during evaluation will be analysed, annotated and added to corpus data, and used to further improve the system performance.

3.1. Wizard-of-Oz set up

As a very first step we analysed some recordings of the famous US game 'What's my line?' whose episodes are freely available on Youtube (`www.youtube.org`). The Wizard script has been designed based on this data. Further, we collected 18 pilot player-Wizard dialogues, for a total duration of 55 minutes comprising 360 system's and user's speaking turns. Pilots' purposes were mainly concerned with the testing of defined Wizard scripts and the design, evaluation and improvement of annotation schemes that are supposed to capture the semantic content of the player's utterances. The collected pilot data is obviously too small to build several types of classifiers, such as ones to classify player's incoming questions, ones to extract answers to these questions, and ones to classify other dialogue acts.

Thus, we collected game/dialogue data in large-scale *Wizard of Oz* experiments. Two English native speakers (1 male and 1 female) were acting as Wizards simulating the system's behaviour. 21 unique subjects were playing the game. The participants were undergraduates of age between 19 and 25, who are expected to be related to our ultimate target audience. 338 dialogues were collected for a total duration of 16 hours comprising about 6.000 speaking turns.

User speech has been transcribed providing word level timings by running the ASR Kaldi (Povey et al., 2011) in ”forced alignment” mode. Transcriptions are stored for each participant and each dialogue separately in a format compliant with TEI standard (ISO, 2006).

The collected and annotated dialogue data serves as a basis for the analysis and modelling of human interactive behaviour, as well as for the design of spoken dialogue system components. In addition to the Dialog Engine components' improvement, the collected speech data is used for ASR development, namely, for generic acoustic and language model adaptation towards the DBOX game scenario.

3.2. Human-machine data collection set up

As mentioned before, based on these dialogues the first version of the dialogue system has been designed. Figure 1 provides the system architecture. The system consists of several modules: Dialogue Engine (DE), Automatic Speech Recognition (ASR), Text-to-Speech synthesizer (TTS), Communication module (COM-Modul), Audio Manager, Game-Client and Game-Server, that are connected and communicate with each other. As soon as any voice activity is detected (VAD), the Game Client with integrated Com-Modul passes this information to the Audio Manager. The spoken input then goes to the ASR module, that prepares the data for the DE in form of recognized tokens hypothesis (possibly more than one per token, e.g. as lattices). The DE consists of four main components, namely interpretation module (IM), dialogue manager (DM), answer extraction module and utterance generation module. The DM takes care of the overall communication between the user and the system. It gets as input a dialogue act representation based on ASR output generated in the IM. In our scenario, questions uttered by the human player play an important role. Questions are classified according to their communicative function (e.g., Propositional, Check, Set and Choice Questions) and semantic content. Semantic content is determined based on Expected Answer Type (EAT) computed from the recognized relations, see Section 4.2. The extracted information is mapped with the EAT and focus word, and the most relevant answer and the strategy how to continue the dialogue are estimated. The DM then passes the system response for speech generation: the DM input is transformed into a dialogue utterance and passed to TTS for speech generation. The output from TTS is returned to the Audio Manager and is subsequently passed to the Game Client to return to the player.

The designed system interface that the players interact with is depicted in Figure 2.

3.3. Non-dialogic data collection

Additionally to the dialogue data, we collected (non-) dialogic data of two types: (1) data containing player's questions that are more realistic than youtube game data and in a larger amount than collected in the pilots; and (2) descriptions of the person to be guessed containing answers to the player's questions.

Dimension	Communicative function	%	Dimension	Communicative function	%	Dimension	Communicative function	%
Task		34.5	Task M.		9.9	SOM		5.3
	SetQuestion	17.0		SetQuestion	1.5		SetQuestion	4.3
	PropositionalQuestion	9.7		PropositionalQuestion	0.5		SetAnswer	4.3
	ChoiceQuestion	3.1		ChoiceQuestion	1.5		Inform	1.4
	CheckQuestions	8.9		CheckQuestions	4.1		Apology	5.7
	Question (unspecified)	1.3		Question (unspecified)	0.8		Thanking	28.6
	SetAnswer	19.2		SetAnswer	3.0		AcceptThanking	8.6
	PropositionalAnswer	8.9		Answer (unspecified)	1.1		InitialGreeting	17.1
	Answer (unspecified)	6.5		Inform	40.9		ReturnGreeting	11.4
	Inform	13.9		Confirm	1.8		InitialGoodbye	5.7
	(Dis-)Agreement	1.3		Agreement	0.5		ReturnGoodbye	2.9
	Confirm	6.1		Suggest	8.3		InitialSelfIntroduction	1.4
	Disconfirm	1.7		Offer	1.8		Congratulation	8.6
	Correction	0.9		Address Offer	1.5			
	Instruct/Request	1.5		Instruct/Request	18.2			
				AddressRequest	12.7			
				Warning	0.5			
Auto F.		16.8	Allo F.		3.6	Turn M.		7.7
	Question (unspecified)	0.9		CheckQuestion	4.2		Turn Take	58.8
	Answer (unspecified)	0.9		Request	4.2		Turn Keep	28.4
	Inform	0.4		Positive AlloFeedback	60.4		Turn Grab	5.9
	Positive AutoFeedback	90.6		Negative AlloFeedback	6.3		Turn Assign	4.9
	Negative AutoFeedback	7.2		Feedback Elicitation	24.9		Turn Accept	2.0
DS		4.1	OCM		2.5	Time M.		14.8
	Opening	27.3		Request	3.0		Stalling	91.4
	Closing	18.2		Self Correction	81.1		Pausing	8.6
	DA announcement	27.3		Signal Speaking Error	15.9			
	Interaction Structuring	27.2						
PCM		0.1	Contact M.		0.7			
	AcceptRequest	100		ContactCheck	44.4			
				ContactIndication	55.6			

Table 1: Distribution of dialogue acts with certain communicative function in DBOX ISO 24617-2 annotated data for each of the addressed dimensions in terms of relative frequency (in %)

For the first type of data, some question data is publicly available, e.g. approx. 5500 questions, annotated according to the scheme defined in (Li and Roth, 2002), are provided by the University of Illinois[6]. However, for our application scenario, this data cannot be used directly, since not all questions are relevant for our domain. We retained only 400 questions for our purposes. Since this obviously constitutes a too small corpus, we generated questions automatically using the tool provided by (Heilman and Smith, 2009) from collected descriptions, and filtered them out manually. In total, 3000 questions were generated. Out of the generated ones, only relevant questions were selected: grammatically broken questions were fixed and repetitions deleted. Additionally, synonyms from WordNet[7] were used to generate different variants of questions for the same class. The final question set consists of 1069 questions.

Answers are retrieved from 100 selected Wikipedia articles in English (71 male persons and 29 female; 51 persons passed away and the others alive) containing 1616 sentences (16 words/sentence on average) and 30.590 tokens (5.817 unique tokens).

4. Semantic annotations

Annotations of two types are performed: (1) dialogue utterances annotations with dialogue act information; and (2) annotations of questions and descriptions with semantic relation information. Both annotation procedures are described in the next two subsections.

ISO 24617-2 dimension	Segmentation (*kappa*)	Coding (*kappa*)
Task	0.88	0.81
AutoFeedback	0.78	0.79
AlloFeedback	0.94	0.95
Turn Management	0.71	0.64
Time Management	0.86	0.86
Discourse Structuring	0.88	0.54
Own Comm. Management	0.55	0.98
Partner Comm. Management	na	na
Social Obligation Management	0.77	1.00
ISO 24617-2 relations		
Functional Dependence	0.88	0.68
Feedback Dependence	0.88	0.88
Rhetorical relations	0.88	0.68

Table 2: Inter-annotator agreement on segmentation and coding per ISO 24617-2 dimension, and on different relation types between dialogue unit as defined in ISO 24617-2.

4.1. Dialogue act annotation: tagset, tool and format

In order to model human dialogue behaviour, it is very common to analyse it in terms of the speaker's intentions. For this, the notion of dialogue act plays a crucial role. Over the years a number of dialogue act annotation schemes has been developed, such as those of the TRAINS project in the US (Allen et al., 1994), the MapTask studies in the UK (Carletta et al., 1996), the Verbmobil project in Germany (Alexandersson et al., 1998). These schemes, however, are not easy to re-use for purposes, or apply to domains, other than the ones they were originally developed for.

In September 2012, the ISO standard 24617-2 "Semantic annotation framework, Part 2: Dialogue acts" has been developed where a comprehensive annotation scheme and markup language DiaML were designed. The ISO 24617-2 standard annotation scheme is a comprehensive, application-independent scheme whose concepts are empirically and theoretically well-motivated, and may be ex-

[6] http://cogcomp.cs.illinois.edu/page/resources/data

[7] http://wordnet.princeton.edu/

RELATION	%	RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACCOMPLISHMENT	4.0%	DURATION	1.8%	LOC_DEATH†	0.8%	PART_IN	3.6%	TIME	14.6%
AGE_OF†	2.1%	EDUCATION_OF†	4.2%	LOC_RESIDENCE†	3.2%	RELIGION†	0.7%	TIME_BIRTH†	2.8%
AWARD	2.5%	EMPLOYEE_OF†	2.2%	MEMBER_OF†	1.8%	SIBLING_OF†	2.3%	TIME_DEATH†	1.0%
CHILD_OF†	3.6%	FOUNDER_OF†	1.2%	NATIONALITY†	3.1%	SPOUSE_OF†	1.9%	TITLE†	14.2%
COLLEAGUE_OF	1.7%	LOC	5.6%	OWNER_OF	1.1%	SUBORDINATE_OF	1.3%		
CREATOR_OF	8.5%	LOC_BIRTH†	5.0%	PARENT_OF†	3.7%	SUPPORTEE_OF	1.1%		

Table 3: List of defined semantic relations and their distribution in data in terms of relative frequency. † means that the relation is adopted from TAC KBP slot filling task.

RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACTIVITY_OF	10.21	LOC_BIRTH	2.34	AGE_OF	3	LOC_DEATH	1.69
AWARD	4.4	LOC_RESIDENCE	1.69	BODY	1.5	MANNER	1.12
CHARGED_FOR	4.21	MEMBER_OF	2.43	CHILD_OF	1.5	NAME	1.87
COLLEAGUE_OF	1.03	NATIONALITY	1.22	CREATOR_OF	6.09	OWNER_OF	1.97
DESCRIPTION	4.12	PARENT_OF	1.31	DURATION	1.31	REASON	1.22
EDUCATION_OF	3.65	RELIGION	2.53	EMPLOYEE_OF	1.59	SIBLING_OF	0.94
ENEMY_OF	1.12	SPOUSE_OF	1.4	FAMILY_OF	1.59	SUPPORTED_BY	0.94
FOUNDER_OF	1.87	TIME	7.96	FRIEND_OF	1.03	TIME_BIRTH	2.06
GENDER	1.69	TIME_DEATH	1.59	LOCATION	4.68	TITLE	11.14

Table 4: Question types in terms of defined semantic relations and their distribution in data (relative frequency in %).

ploited for constructing annotated dialogue corpora. In DBOX, we used this ISO dialogue act annotation scheme. ISO 24617-2 is a highly multidimensional scheme supporting multifunctionality, since it offers the possibility to assign multiple dialogue act tags to one dialogue segment. The ISO 24617-2 taxonomy of communicative functions distinguishes 9 dimensions: addressing information about a certain (*Task*); the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*). For DBOX purposes, we considered 2 additional dimensions (11 in total): *Contact Management*, which is non-core optional in ISO24617-2, since DBOX games are not face-to-face dialogues managing the contact is an important aspect in such types of dialogues; and *Task Management* for dialogue utterances concerned with game rules. There are 41 dimension-specific and 26 general-purpose communicative functions. Not all of ISO functional tags occur in our data. On the other hand, we introduced 2 additional dimension-specific functions and 1 general-purpose function that are not included in ISO 26417-2, however, defined in DIT^{++} (Bunt, 1999) that ISO taxonomy is based on:

- *Dialogue Act Announcement*, where the speaker makes explicit what kind of dialogue act he/she is going to perform next;
- *Preclosing*, where the speaker indicates that he/she plans to end the current dialogue shortly;
- *Threat*, where the speaker states his committment to perform the action in the manner or with the frequency, described in the semantic content; speaker believes the action to be harmful for the addressee

We also introduced 1 dimension-specific function that is not present in above mentioned taxonomies, but frequently occurs in DBOX data, namely, *Congratulation*, where the speaker wants the addressee to know that the action the addressee performed recently was successful and/or of good fortune for the addressee. The standard allows adding domain-specific communicative functions provided they are observed in data, relevant for adequate coverage, and human (and machine) recognizable. In DBOX dialogues *Congratulations* are mainly performed by the system (or Wizard) when the player guessed the person's identity correctly and thereby won the game.

Annotations are of stand-off type and performed with the ANVIL tool[8] using the specification designed for ISO 24617-2[9] which allows us to convert data into Dialogue Act Markup Language – DiAML (see Bunt et al., 2012). The ANVIL tool allows annotations in multiple tiers so that for each participant we specified a speech tier and several tiers for each dimension. Dialogue act annotations are saved both in `.anvil` and `.diaml` formats. All 18 pilot dialogues, and out of the collected 338 dialogues 60 dialogues are annotated with dialogue act information. Table 1 presents the distribution of the most frequent dialogue acts in the collected corpora per the addressed dimension.

In order to assess the reliability of the selected dialogue act tagset on our data, we measured the inter-annotator agreement in terms of the standard Kappa statistic (Cohen, 1960). Pilot dialogues were annotated by one expert and one trained annotator. The trained annotator received ap-

[8] For more information about the tool visit: `http://www.anvil-software.org/`

[9] The specification is available at `http://www.anvil-software.org/data/diaml-spec-v0.5.xml`

Type	Content	Format	Comment
Metadata	participants (id native language sex age at collection) list of system's characters	xml	
	corpus description	xml	
Signals	sound recordings	wav 16-bit	1 channel per speaker
Transcriptions	tokens(id, start, end, string)	TEI compliant	semi-automatic
DA annotations	dialogue act (sender, communicative function dimension qualifier functionalDependenceRelation feedbackDependenceRelation) rhetoricalLinks	Anvil and DiAML	manual
Persons descriptions	tokenized lemmatized POS-tagged	csv csv csv	Stanford Core NLP
	Chunking	csv	OpenNLP
	NE tagged	csv csv csv	Stanford NER Illinois NER Saarland NER
	Semantic relations	csv	manual
Question corpus	tokenized lemmatized POS-tagged	csv csv csv	Stanford Core NLP
	Chunking	csv	OpenNLP
	NE tagged	csv csv csv	Stanford NER Illinois NER Saarland NER
	Semantic relations	csv	manual

Table 5: DBOX corpus overview.

proximately 5 hours annotation training divided into three sessions. Table 2 presents the kappa results for each ISO 24617-2 dimension separately. Agreement was measured on both segmentation and classification of dialogue acts. Additionally, agreement on three types of relations between identified dialogue units was assessed. The obtained kappa scores were interpreted as annotators having reached a good agreement.

4.2. Semantic content: relations

The set of the 1400 most frequently asked questions, and 100 annotated Wikipedia descriptions of famous persons, that form the main game content, were semantically annotated. In order to find the answer to certain questions, semantic role information is often used. A semantic role is a relational notion (between an event and its participant) and describes the way a participant plays in an event or state, as described mostly by a verb, typically providing answers to questions such as ”who” did ”what” to ”whom,” and ”when,” ”where,” ”why,” and ”how.” For our purposes, however, we are also interested in relations between participants. Indeed, along with semantic roles, relations between participants are relevant for our domain, e.g. the relation between Agent and Co-Agent involved in a ‘work’ event may be a COLLEAGUE_OF relation.

From the Wizard-of-Oz experiments we observed that players tend to ask similar questions about gender, place and time of birth or death, profession and titles, achievements and awards, marriage, children, etc. After data analysis, there were 59 semantic relations defined, where 17 have been adopted from TAC KBP 2013 Slot Filling[10]. TAC relations are mainly defined for relations between Named Entities (NE) such as persons and organizations, while our proposed set incorporates temporal event markers like TIME (which may be further subdivided into Initial and Final Time), DURATION and FREQUENCY; captures PURPOSE and CAUSE relations between events; and introduces the event MANNER marker. Additionally, our set captures 12 more relations between entities such as ACCOMPLISHMENT, AWARD, CREATOR_OF, COLLEAGUE_OF, OWNER_OF, SUPPORTER_OF, etc., and are not restricted to relations between NEs.

Table 3 gives an overview of the most frequently occurring relations (out of 3988 identified) in the persons’ description data, and Table 4 presents the relative frequencies of questions types based on the identified semantic relation.

All the questions in the set, as well as the persons descriptions, were tokenized and lemmatized. Subsequently, Stanford Core NLP[11] has been used for extracting the POS and NE information. In addition, openNLP chunker, and NER such as Stanford NER, Illinois NER and Saarland NER were applied. All these annotations are stand-off annotations and will be provided with the DBOX dialogue corpus. To assess the reliability of the defined semantic relation tagset, the inter-annotator agreement was measured . For this purpose, 10 randomly selected descriptions were annotated by two trained annotators. The obtained *kappa* scores were interpreted as annotators having reached good agreement (averaged for all labels, kappa = 0.76).

5. Corpus overview and future work

The DBOX dialogue corpus has required substantial investment. We expect it to have a great impact on the rest of the project. The DBOX project consortium will continue to maintain the corpus and to take an interest in its growth, e.g., expand to other languages. In the future, the well documented DBOX corpus will be available for the research

[10] http://www.nist.gov/tac/2013/KBP/

[11] http://nlp.stanford.edu/software/corenlp.shtml

community. The human-human DBOX English corpus will be made available for research purposes in autumn 2014. Table 5 provides details on the future corpus release. The human-computer data collection will start after most of the Wizard-of-Oz data is processed (in summer 2014). The global DBOX architecture and corpus collection strategy that have been applied to quiz game and discussed in this paper could be applied to other domains and languages.

6. References

Allen, J. et al. 1994 The TRAINS Project: a case study in building a conversational planning agent. *TRAINS Technical Note 94-3*. University of Rochester.

Alexandersson, J., et al. 1998 Dialogue acts in Verbmobil-2. Second edition. *Report 226*. DFKI Saarbrücken, University of Stuttgart; TU Berlin; University of Saarland.

Bunt, H. 1999 Dynamic interpretation and dialogue theory. In M. Taylor, F. Neel, and D. Bouwhuis (eds.), *The structure of multimodal dialogue II*, Amsterdam: Benjamins, pp. 139–166.

Bunt,H., Kipp, M., and Petukhova, V. 2012 Using DiAML and ANVIL for multimodal dialogue annotation. In Proceedings 8th International Conference on Language Resources and Evaluation, Istanbul, May 2012. ELRA, Paris.

Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. 1996 *HCRC Dialogue Structure Coding Manual*. Human Communication Research Centre HCRC TR-82, University of Edinburgh, Scotland.

Cohen, J. 1960 A coefficient of agreement for nominal scales. *Education and Psychological Measurement* 20:37.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. 1993 Wizard of Oz studies - why and how. In Proceedings of the International Workshop on Intelligent User Interfaces, pp. 193–200.

Heilman, M. 2011 Automatic Factual Question Generation from Text. PhD thesis, Carnegie Mellon University, US.

ISO. 2006 TEI-ISO 24610-1:2006 Language resource management: Feature structures, Part 1: Feature structure representation.

ISO. 2012 Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2, Geneva, ISO, Geneva.

Li, X. and Roth, D. 2002 Learning question classifiers. In: Proceedings of the COLING '02, Taipei, Taiwan, Association for Computational Linguistics, pages 1–7.

Paek, T. 2006 Reinforcement Learning for Spoken Dialogue Systems:Comparing Strengths and Weaknesses for Practical Deployment. In Proceedings of the Interspeech-06 Workshop on 'Dialogue on Dialogues - Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems'.

Povey, D. et al. 2011 The Kaldi Speech Recognition Toolkit. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding.

Appendix: dialogue example

S: Hello
P: Hello
S: Good afternoon almost evening
S: What is your name
P: My name is James
S: Hello James it's nice to meet you
P: Nice to meet you
S: How are you doing today?
P: Good, thank you
S: Alright
S: Today we are going to play a game and here are the rules
S: I'm a very famous person and you need to guess my name you can ask whatever questions you want of me except for my name directly
S: You have at most ten questions and then you get to guess my name exactly once
S: So you can ask whatever questions you want but then if you want to guess my name you only get one try
S: If you get my name correct you win if you get my name incorrect or choose to pass then you lose and then we'll move on to the next round
S: Do you understand and are comfortable with the rules?
P: Yeah yeah
P: So the name is kind of a famous person
P: Okay
P: I'm not sure how good am I in this area
S: Yes
S: I am a famous person and I am male
P: Okay okay good
S: Alright
S: And what is your first question?
P: What is the first question
P: What do you do?
S: I am a leader
P: A leader
P: What is your nationality?
S: I am American
P: Are you alive?
S: I am not alive
P: Are you leading a company?
S: I am not leading a company
P: okay
P: You're not a company leader
P: When are you born?
S: I was born on february twenty second seventeen thirty two
P: Seventeen thirty two
P: Ok
P: Eehm
P: Are a politician?
S: I am a politician
P: Okay
P: So then it is not my area but I will try to guess
P: When were you in the government?
S: Uhm
S: Let's see
S: I retired from the presidency in seventeen ninety seven
P: Ninety seven
P: George Washington S: Is that your final guess? P: Yes, Washington
S: Very good, excellent job!
S: Congratulations!