BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction

Jan Frederik Maas*, **Britta Wrede***

*Bielefeld University
Technical Faculty, Applied Computer Science Group
Universitätsstr. 25
33615 Bielefeld
{jmaas, bwrede}@techfak.uni-bielefeld.de

Abstract

Our research is concerned with the development of robotic systems which can support people in household environments, such as taking care of elderly people. A central goal of our research consists in creating robot systems which are able to learn and communicate about a given environment without the need of a specially trained user. For the communication with such users it is necessary that the robot is able to communicate multimodally, which especially includes the ability to communicate in natural language. We believe that the ability to communicate naturally in multimodal communication must be supported by the ability to access contextual information, with topical knowledge being an important aspect of this knowledge. Therefore, we currently develop a topic tracking system for situated human-robot communication on our robot systems. This paper describes the BITT (Bielefeld Topic Tracking) corpus which we built in order to develop and evaluate our system. The corpus consists of human-robot communication sequences about a home-like environment, delivering access to the information sources a multimodal topic tracking system requires.

1. Introduction

Our research takes place within the "Cognitive Robot Companion" (COGNIRON[1]) project. The project is concerned with the development of robotic systems which can support people in household environments, for example by taking care of elderly people, or performing everyday tasks.
A central milestone of the project is to create robot systems which are able to learn and communicate about a given environment without the need of a specially trained user. Thus, our research is focused on building a robotic system being capable of multimodal communication, especially natural language. The communication capabilities of the robot system should be as natural as possible, because constrained communicating systems – for example dialogue systems not capable of proper anaphora resolution – put additional workload on a non-specialist communication partner.
Based on these considerations, we designate human-robot interaction as "natural" when

1. the used modalities (language, gestures, etc.) are the same as for face-to-face human-human communication
2. the human does not have to learn how to communicate with the robot, but can apply his or her knowledge of human-human-communication.

In natural communication, knowledge about the *context* of the communication is necessary. One part of contextual information is the knowledge about the current *topic*, which can be used for many communicational tasks, e.g. anaphora resolution, managing background knowledge, etc.
Since a household robot needs to be able to adapt to new and changing situations ("open-endedness"), it is not sufficient to use predefined topics. Thus, we decided to build a robot being capable of learning new topics by analysing dialogues online. To be able to develop and evaluate such a system, we built a corpus containing the relevant information the system could acquire during a communication.
In contrast to the corpus created by (Green et al., 2006) which focuses on interactive aspects such as communicative problems and spatial relations during communication, the BITT corpus is strongly focused on capturing higher dialogue structures (i.e., topics) emerging during human-robot interaction. The corpus is designed to deliver preprocessed data for topic tracking algorithms[2], facilitating the development and evaluation of such algorithms without an online robot system.

2. Corpus design

In order to get the information a mobile robot system could acquire during a communication, we decided to record the corpus from the robot's perspective. To be useful for our research, the corpus had to contain natural – not artificially constrained – communication sequences. Additionally, the corpus should contain rich situated topic information.
The setting of the corpus is a so-called Home-Tour scenario, during which a subject introduces a robot to a household-like environment. The advantages of such a scenario for our task are:

1. Home-Tour scenarios resemble basic communication situations for a household robot.
2. The topical structure of the communication is mainly controlled by the subject and not by the robot or both communication partners.
3. The topical structure of the communication is likely to reflect the physical structure of the experimental setting, making it possible to enforce a rich topic structure by the design of the setting.

[1](Cogniron, homepage)

[2]for example, cf. (Allan, 2002)

2.1. Hardware

For the recording of the corpus we used the mobile robot BIRON[3] as a platform (cf. Fig.1).

Fig.1 - BIRON

BIRON is a modified ActivMediaPeopleBot. It is able to detect people, communicate by spoken language, understand simple pointing gestures and detect specified objects. It has the ability to show several facial expressions on its display, thus communicating different states of attention and reporting communication problems in a simple way.
Its sensory equipment consists of:

- stereo microphones
- a pan-tilt camera for face tracking
- a laser range scanner
- a stereo camera

We recorded each sensory source except for the laser range scanner, which is only used for person detection tasks, but not for higher dialogue functions. The data gathered from the laser range scanner was used for person tracking during the experiments, though.
Additionally, we recorded the experiments with an external camcorder connected to a headset microphone. This facilitated the manual postprocessing – especially the transcription – of the corpus.

2.2. Experimental design

We invited 29 people to show a specially prepared room to the robot BIRON. The subjects were told that the robot needed the information to introduce the room to seven year old children afterwards. We chose this scenario in order to bias the subjects to elaborate more on the contents of the room. In several cases the people directly instructed BIRON to instruct the "children" about dangerous objects or things to play with, indicating that this approach was successful.

[3]cf. (Haasch et al., 2004)

2.2.1. Setting

As can be seen on Fig.2 and Fig.3, the room contained several topical areas.

Fig.2 - Part of the setting

Examples for topical areas are a kitchenette, a working place, a place to have a cup of tea, etc. In most cases, the topical areas were adopted by the subjects during the description of the room. Only in a few cases, topics spanning the topical areas were developed, for example a topic concerning all the plants in the experimenental room.

2.2.2. Robot behaviour

During the experiments, the attention system of BIRON was activated. Thus, it simulated attention by movements of the pan-tilt camera and rotation of the base, tracking the subjects' bodies and faces (cf. Fig.3). The display showed different facial expressions depending on the attention states, e.g., listening or waiting.

Fig.3 - Human-robot interaction example

In order to not restrict natural communication of the subjects by a restricted dialogue system or speech recognition errors, we deactivated the verbal communication capabilities of the robot. This way it behaved only as a listener, but simulated attention by the above mentioned robot reactions. Although the missing verbal feedback of the robot could be estimated as a drawback, we decided not to carry out a Wizard-of-Oz style experiment in order to avoid subject biasing. Also, it is not absolutely clear what tasks future

interaction robot systems will be able to accomplish, so the definition of a simulation of the robot's (i.e., the wizard's) capabilities would have been somewhat arbitrary.
The resulting monologues were an ideal starting point for our research, because they contained free, unconstrained speech, only user-initiated topic shifts and only few disturbances because of robot errors.

3. Postprocessing

Based on hypotheses of cues relevant for topic tracking, we were especially interested in the following information:

1. Pauses (indication of topic shifts)
2. F0 information (indication of topic shifts).
3. Lemmatised spoken language (indication of topics)
4. References to objects or groups of objects by language, gestures or both (indication of topics in situated communication)
5. For the evaluation task we additionally needed information about topics annotated by humans.

Note that except for 5., each of these types of information are or will be available to BIRON during a communication. However, for the corpus we simulated most of the processing steps by manual annotations, to reduce the error rate and get an optimal base for Topic Tracking.

4. Annotation

According to the stated requirements, we decided to include all the data – except for the F0 information – in one single XML-format. For each monologue, phases of spoken language were detected by the automatic voice activity detection[4] (VAD) we use on BIRON. The monologues were thoroughly transcribed and afterwards enriched by information on multimodal references to objects or object groups, based on gestures, reference by language or both. However, note that no explicit gesture annotation was performed. Only objects that were referenced by deictic gestures were marked in the corpus.

4.1. Time and pause information

The main elements of the corpus are communication segments, i.e., utterances. Each utterance is a continuously spoken part of language which is assumed to bear no topic shifts, i.e. a single – or no – topic. As mentioned above, the utterances were detected using an energy based VAD system. Each utterance bears a start and an end attribute in the corpus files. This way, the start and end attributes of consecutive utterances define speaker pauses. For example:

<utterance start="35.11" end="36.39"> (...) ***and a lot of things*** *(...)</utterance>*
<utterance start="36.63" end="41.22"> (...) ***this desk for example*** *(...)</utterance>*

Example 1. - Time Information

We found that pause length is a useful indicator of topic shifts in the corpus.

4.2. F0 Information

The F0-information was recorded in time-aligned text files generated by PRAAT, cf. (Boersma and Weenik, 2004). We decided to analyse the data gained by the headset microphone instead of the robot microphones because of sound quality reasons.

4.3. Transcription

As mentioned above, the utterances of the subjects were graphemically transcribed. We used a simple annotation scheme to mark aborted utterances or unusually pronounced words. Noise, such as breathing of the subjects, was annotated for speech recognition purposes, but this information was deleted from the topically annotated version of the corpus.
A fixed list of hesitations was defined and used.
For most topic tracking and information retrieval tasks, lemmatisation or stemming is a necessary preprocessing step[5]. In the BITT corpus each utterance is specified as well in a lemmatised as in a not lemmatised form. The given lemmatisation of the BITT-corpus was performed with TreeTagger[6].

4.4. Reference solution

One of the key aspects of the corpus is the annotation of object references. This information can be used to disambiguate object references from different topics which bear the same verbal description, e.g., "the plant", but refer to different plants in different contexts.

4.4.1. Object resolution

The annotation differentiates between references to object groups and single objects. Objects were given an ID. References to objects outside the room were not annotated.
Verbal object references were annotated as well as gestural or combined references, although no difference in the annotation scheme was made. In case of verbal references, the referring NP was annotated (cf. Example 2.1). In case of purely gestural object references, the part-of-speech tags were left empty (cf. Example 2.2)

there is *<object><reference oid="refrigerator_01"/> <pos>* ***a fridge*** *</pos> </object>(...)*

Example 2.1 - Verbal object reference resolution

look *<object><reference oid="plant_02"/> <pos/> </object>*

Example 2.2 - Gestural object reference resolution

Purely gestural object references were very rare in the corpus.

4.4.2. Object group resolution

We decided to differentiate between three types of object groups. This distinction was made in order to reduce the object space but to be able to consider references on object groups as well. The three types of object groups are:

[4]The VAD system is part of the ESMERALDA toolbox, see (Fink, 1999)

[5]cf. (van Rijsbergen, 2005)

[6]see (Schmid, 1994)

1. Groups consisting of items which are almost always[7] referred to as one single group (e.g., a set of chocolate bars)

2. Abstract groups (e.g., all humans)

3. Groups consisting of individual items which are sometimes referred to as single items and sometimes are mentioned as a part of a group (e.g., a group of plants and the subject refers to the single plants afterwards).

Groups of type (i) and (ii) were treated as single objects, i.e., they were assigned an id:

(...) ***if they want to have*** *<objects><group oid= "g_sweets_01" /> <pos>**them**</pos></objects> (...)*

Example 3.1 - Object group reference resolution

Groups of type (iii) were annotated by specifying the ids of the objects contained:

(...) ***here two*** *<objects><group oid=""/> <member oid="plant_05"/><member oid="plant_01"/> <pos>**flowers**</pos></objects>* (...)

Example 3.2 - Object group reference resolution

4.5. Topic annotation

Three trained annotators annotated the topic for each utterance. Each annotator had to decide on a non-hierarchical set of topics before beginning the annotation work, but after familiarising him- or herself with the data. Thus, mainly global topics, i.e. topics occurring in more than one monologue, were annotated. We intended to annotate mainly global topics, because for our applications, the detection of local topics is only of limited use: Global topics bear information about recurring events, tasks etc., while local topics do not.
Each annotator assigned one or no topic to each communication segment – communication segments without any topical indication, e.g. "what next", were intentionally not annotated. This way, the complete corpus was annotated by each annotator, yielding three different topic annotations.

5. Statistics

The BITT corpus consists of 29 transcribed and annotated monologues of 24 female and 5 male subjects, as well as the recorded data from the mentioned sensory sources. The total recording time was 320 minutes, with a mean of about 11 minutes per monologue. The language of the experiments was English, although most of the subjects[8] were non-native speakers.
2620 references to single items were annotated, as well as 1419 references to groups of objects. The corpus consists of 11209 communication segments. On average 4900 communication segments were annotated with a topic by each annotator.

[7] i.e., in the context of the corpus

[8] There were 2 native and 27 non-native subjects. Each of the subjects estimated his/her English speaking capability as good or better.

6. Outlook and Availability

The BITT corpus is currently used in the development and evaluation of the Topic Tracking software of BIRON. The parts of the corpus which are freely (no fee) available contain:

1. The annotated transcriptions as described above.

2. The time aligned pitch analysis files.

Other parts of the corpus are protected by German data protection law and can only be made available under special circumstances.
If you are interested in using the BITT corpus for research purposes, please contact us by one of the email addresses given in the header.

7. Acknowledgments

This research is supported by the Graduate Program "Task Oriented Communication", funded by the German Research Foundation, and by the COGNIRON-project (FP6-IST-002020), supported by the European Union.

8. References

James Allan (ed.) 2002. *Topic Detection and Tracking*. Kluwer Academic Publishers, Norwell, Massachusetts.

Paul Boersma and David Weenink 2004. Praat: doing phonetics by computer (Version 4.2.17) [Computer program]. URL: http://www.praat.org/

COGNIRON - The Cognitive Robot Companion – Project homepage. URL: http://www.cogniron.org/

A. Green, H. Hüttenrauch, E. A. Topp and K.S. Eklundh 2006. *Developing a Contextualized Multimodal Corpus for Human-Robot Interaction*. International Conference on Language Resources and Evaluation (LREC), Genua.

A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede and G. Sagerer. 2004. *BIRON – The Bielefeld Robot Companion*. E. Prassler, G. Lawitzky, P. Fiorini and M. Hägele (ed.): Proc. Int. Workshop on Advances in Service Robotics, pp. 27–32.

1999. G. A. Fink: *Developing HMM-based recognizers with ESMERALDA*. V. Matousek, P. Mautner, J. Ocelkov, and P. Sojka (ed.): Lecture Notes in Artificial Intelligence, volume 1692, pp. 229–234, Berlin Heidelberg. Springer.

2005. C.J. van Rijsbergen: *Information Retrieval*. URL: http://www.dcs.gla.ac.uk/Keith/Preface.html, 2nd edition.

1994. H. Schmid: *Probabilistic Part-of-Speech Tagging Using Decision Trees*. International Conference on New Methods in Language Processing.