# Towards a Multimodal Topic Tracking System for a Mobile Robot

*Jan F. Maas, Britta Wrede, Gerhard Sagerer*

Applied Computer Science Group
Faculty of Technology, Bielefeld University, Germany
{jmaas,bwrede,sagerer}@techfak.uni-bielefeld.de

## Abstract

Topics in situated and task oriented communication depend heavily on the given, often changing environment, making the detection of predetermined topics in many cases useless. Detection of non-predefined topics can enhance Human-Robot-Interaction (HRI) in a variety of ways, though.

In this paper we propose a way to dynamically determine topics during Human-Robot-Communication using well established techniques such as Latent Semantic Analysis (LSA). The procedure is based on multimodal cues, supporting the view that topics are not simply a property of spoken or written language, but of multimodal situated communication. An online version of the topic detection system has been developed and is currently being tested on our mobile robot BIRON. To demonstrate the feasibility of our approach, we present the results of an evaluation of our system on the BITT corpus.

**Index Terms**: topic tracking, multimodal dialogue, human robot interaction.

## 1. Introduction

During the last years, topic tracking has been used to support many different tasks in natural language processing. For example, topic tracking techniques have been developed in the context of information retrieval [1] to make relevant information from written (e.g., newspaper) corpora accessible. Also, topic detection and tracking techniques have been developed for broadcast news [2] or spoken language databases [3], thus allowing topic sensitive search on broadcast news databases or building intelligent radios, which scan for selected information.

In our work we want to make the benefits of TDT in situated, task-oriented Human-robot-interaction (HRI) accessible. The goal is the design of a system which allows a robot to track topics during communication. The knowledge about the relevant topic can support several tasks of the robot, such as speech recognition, speech understanding or social interaction. For example, speech recognition can be enhanced by adjusting the language model according to the current topic, while speech understanding can be enhanced by resolving underspecified references: The sentences "My shoes are dirty. Fetch me a brush" do not specify the kind of brush. Topical knowledge can help determining that a shoe brush is asked for.

Generally speaking the TDT information can be used for adaptation of the robot to a given environment, allowing "soft" adaptation rather than learning "hard" facts, i.e., propositions.

### 1.1. Scenario

The research is embedded in the framework of the COGNIRON-project [4]. The main goal of the project is to study perceptual, representational, reasoning and learning capabilities of embodied robots in human centred environments, e.g., private homes.

Within this framework, several scenarios were defined to obtain simplified access to the complex situations that can occur. Our TDT research is focused on the so called "home tour" scenario where a robot is introduced to an unknown private home environment. The robot should be able to communicate by means of spoken language but also be able to make use of multimodal information as well. It should be able to follow users, ask questions and learn about the environment.

#### *1.1.1. Comparison to other scenarios*

In comparison to other application areas of Topic Spotting or TDT, several differences are obvious. A difference which makes topic detection in situated communication difficult concerns the size of the entities which are to be classified: A newspaper article usually has more than hundred words, containing many features allowing a classification. In HRI it is possible to have multiple topic changes during a single dialogue turn.

Another challenge is the necessity to detect *dynamic* (not predefined) topics, emerging through the structure of the environment. E.g., it is not useful to predefine topics such as "living room" for there may be none, whereas a billiard room may exist which is very unlikely in most homes and thus to appear in a list of predefined topics. These circumstances make solely word-based dynamic TDT in situated HRI nearly impossible. To cope with the difficulties we propose the use of multimodal information such as gestures, object references, identification of communication partners, etc., which will be available to our robot system.

### 1.2. Related work

While the detection of topics in situated, multimodal HRI is a new area of research, our method makes use of former successful approaches of topic detection. The structure of the system roughly follows the tasks defined in the Topic Detection and Tracking project [2]. A related approach of dynamic topic detection on spoken language was developed by [3]. The capability of Semiotic Spaces – which are the basic technique of our topic tracking approach – to cope with additional, dialogue related information was shown in [5]. Our approach of segmentation of spoken language into topics made use of the results in [6].

While most of these approaches have been proved to work on large non-situated corpora, it is our goal to develop a strategy for topic tracking in situated communication with few training data supported by multimodal information.

## 2. Topic Tracking Approach

Before we explain the method underlying our approach, we want to define the notion "topic". Intuitively, a topic is "what a text or discourse segment is about". It is possible to have several names for a topic, e.g., "washing clothes" or "how to do the laundry" etc. ([7], p 72), therefore we do not indicate topics by speaking names but by sets of signs which are thematically related (e.g., "washing machine", "washing powder", "water", "tap" etc.).

In our approach we define a **topic** $\tau$ as an abstract property of communication segments. A **communication segment** is a collection of signs perceived by a robot during a certain period of time. For example: A human shows the task of dish-washing to a robot. The words uttered by the human, but also the objects involved in the interaction and possibly more signs build a communication segment.

A communication segment is assumed to bear exactly one topic, although several consecutive communication segments may have the same topic. Merging such communication segments would result in another, larger communication segment.

A perceived sign, which is an element of a certain communication segment, may indicate the topic of the communication segment. For example, the word "fridge" uttered during a communication segment may indicate a topic centred around a kitchen. Also, the gestural reference to the fridge can do the same.

### 2.1. General Idea

It is our goal to dynamically build and detect topics emerging during a human-robot-interaction. To reach this goal, we gather information about signs commonly co-occurring during communication segments, which are derived from a continuous communication during a segmentation process. Signs which commonly co-occur are assumed to belong to the same topic. This way, the signs of former communication segments can - if encountered again - be used to recognise the current topic. The information gathered about this topic so far can then be used for different tasks.

In our approach, a topic is represented by the set of all topic indicating signs (e.g., function words – like "and" – which indicate no topic are not part of this set) and the indication strength of the sign for the topic (see below). This way, a simple weighted unigram classification can be carried out to detect the topic of newly encountered signs, i.e. communication segments.

### 2.2. Why using Semiotic Spaces?

The idea of representing semantic relationships by utilising co-occurrence information is the central idea of Semiotic Spaces [8]. In information retrieval research Semiotic Spaces have been successfully used to build up groups of texts or words which are similar or belong to the same topic [9]. Semiotic Spaces are not restricted to words or a single source of information, but can be applied to data from mixed sources of information as well [5]. Additionally, Semiotic Spaces can be applied to very small training sets, thus allowing for topic tracking in very short communications. Further, they are robust against dialogue parsing errors, since they rely on symbol-based cues. For these reasons, we decided to use Semiotic Spaces as the core technique for our research.

### 2.3. Types of Semiotic Spaces

Semiotic Spaces are in most cases high-dimensional vector spaces into which texts or words are projected. In these spaces, the distances of the word or text vectors are a measure of similarity, similar entities being close to each other. The more often signs co-occur, the smaller the distance between the signs will become.

In text technology several kinds of Semiotic Spaces exist [8]. The Semiotic Spaces differ in their power to represent semantic associations and their computational speed, but additionally in the need for specifying the number of topics beforehand. For this reason, we did not use Support Vector Machines (SVM) or Probabilistic Latent Semantic Analysis (PLSA), because the highly dynamic household environments do not allow to specify the number of topics in advance.

For our experiments we employed three Semiotic Spaces: The simple vector space model [10], Fuzzy Semantics [11] and LSA [9]. In the simple vector space model, each sign is represented by a vector $v = (d_1, ...d_n)$, with $d_n$ being the entropy-weighted frequency of the sign in the c'th communication segment of the training set. Fuzzy Semantics spaces are normalised simple vector spaces, while LSA employs dimensionality reduction on the data to reduce the impact of "noise".

### 2.4. Procedure

The topic tracking procedure consists of five steps in order to compute the indication strength for each topic indicating sign: Segmentation, preprocessing, semantic association, clustering of the Semiotic Space and estimation of support strength. These steps are computed on training data. Segmentation and preprocessing are also preparatory steps for the Topic Tracking on new communication segments. Depending on the hardware and algorithms used, each new communication segment can be employed for training after its classification, enabling topic learning during the communication.

**1. Segmentation:** The goal of the segmentation is to split a communication process into communication segments, which ideally have exactly one topic. During the segmentation phase, signs from different modalities are merged in a "bag of signs" which constitute the communication segment. The segmentation process is the main information source for the unsupervised learning task, which takes place during the calculation of the Semiotic Space and clustering. While segmentation is an important issue for our topic tracking approach, it heavily depends on the kind of communication. For our system, we use several multimodal cues such as movement of the user, eye (camera) contact and dialogue structure.

Consider that two different segmentation processes take place for training and tracking: While the segmentation process for training results in long communication segments to gather as much co-occurrence information as possible, the segmentation process for the tracking delivers short pieces of continuous communication, e.g., utterances. This is done to allow real-time tracking. To avoid confusion, we will call the segments which result from the first method "communication segments", and the segments for tracking "chunks". Further, the notion "segmentation" will only be applied to the process of creating communication segments. Chunks are gained by an automatic voice activity detection both for the online system and the offline experiments.

**2. Preprocessing:** Preprocessing is concerned with deleting words that are unlikely to indicate topics, e.g., function words, not conventionalised gestures, etc. Also, signs may be transformed into more abstract signs when they have the same relevance for a topic (e.g., by lemmatisation). In our system, a stoplist of function words was manually created. Lemmatisation is also done by a manually edited lookup table, although more elaborated algo-

rithms exist and were utilised for the offline experiments (see below).

**3. Semiotic Space:** This step is needed to compute the distances of signs to each other, building the necessary foundation for a topic clustering. Distances can be computed with several standard distance measures such as cosine, Euclidean distance, etc. In our system, we usually apply correlation coefficients.

**4. Clustering:** The Semiotic Space only represents thematical distances between signs. To gather information about interrelated signs, i.e., topics, a clustering algorithm is applied. For our purposes, we use average-linkage agglomerative clustering [12], since it is fast and allows for dynamic segmentation by specifying a clustering limit. During the experiments, the algorithm was constrained to detect 10 clusters or less – slightly more than in the manual annotation – to support comparability.

**5. Classification:** Preceding classification, the average distance of each sign to each sign within a topic cluster is computed. To determine the topic of a chunk, we simply sum the distances of the signs in the chunk for each topic, thus implementing a weighted unigram classification. The topic with the highest value is the detected topic. Should no topic be detected because of no topic indicating signs, the last detected topic is registered.

Consider that in the online system, steps 3 and 4 are continually repeated during communication, thus updating the system. For offline evaluation, topic clusters based on a training set were created and not modified during tracking (see below).

# 3. Corpus

Evaluation of our topic tracking system was performed on the BITT-corpus [13], which was developed for evaluation of topic tracking on mobile robots. The corpus consists of 29 human-robot interactions in a scenario similar to the home tour. The subjects introduced a room they had familiarised themselves with to the robot BIRON [14]. The room was specially prepared to contain a lot of items which topically belong together, for example a place to play games, a kitchenette, a working place, etc. Thus we hoped, different topics related to the different areas would occur naturally during the communication.

BIRON is a modified Pioneer PeopleBot from ActivMedia. It was able to turn the base during the experiment, to watch the instructor and to track the face of the user with a pan-tilt camera. Thus, the robot simulated attention. No more robot reactions took place in order to minimise communication problems, allowing the people to communicate most naturally and without special instructions. Because of the small number of constraints the resulting data was highly heterogeneous: People used different communication strategies, different detail levels, etc. The duration of the experiments varied from 3 minutes up to over 30 minutes, with about 9.5 minutes average. The interactions were recorded and manually transcribed.

**Annotation** After the transcription, object references were annotated, thus simulating the capability of our robot to detect objects which were verbally and/or gesturally referred to. Each object was given an ID. In some cases, object groups were marked, e.g., a set of chocolate bars which were usually referred to as a set were given a single ID. Consider that almost no gestures referring to objects occurred without accompanying speech, although the gestures were in many cases necessary to resolve the reference.

The signs (words and object references) were grouped in short intervals of continuous speech – the chunks which are classified during tracking – by BIRON's voice activity detection. Time information (start and end) was added to each chunk.

To facilitate topic tracking evaluation, three annotators noted a topic for each chunk. The annotators decided on the topics for themselves before annotation and after making themselves familiar with the corpus. This way, dynamic topic tracking on the corpus was performed manually.

# 4. Evaluation

For evaluation, each of the 29 monologues of the BITT corpus was tracked by our system. The training set for each monologue consisted of the 28 other monologues. Lemmatisation of the words of the corpus was performed with the TreeTagger [15].

To create communication segments for building the training sets, subsequent chunks bearing the same manually annotated topic from the topic annotation of the BITT corpus were merged.

## 4.1. Evaluation metric

We compared the results with the three manual topic annotations of the BITT corpus. However, a direct comparison is not possible because of two reasons: First, the number of automatically and manually detected topic may change from communication to communication. Second, a decision process is needed for determining which dynamically detected topic represents which manually detected topic.

To cope with these problems, we developed the following evaluation metric:

1. The chunks which were automatically classified as belonging to a single topic were determined.
2. For these chunks, the quantity of chunks belonging to each manually determined topic was determined.
3. The chunks bearing the manually detected topic with the highest count were considered as correct, the chunks bearing other manually detected topics were considered as false.

This process was repeated for each automatically detected topic, the results were summed up. With nearly 4900 manually annotated chunks for each of the three annotators, each evaluation is based on about 14.500 single values.

The topic tracking system was biased to detect about as many topics as were determined by the human annotators by specifying a maximum number of topics to be detected.

## 4.2. Results

We now want to discuss the results from the offline evaluation process. Since we believe that for topic tracking on situated HRI a multimodal approach is needed, we varied the available sign types for training. The system was tested with words (Table 1), with verbal object references replaced by object IDs (Table 2) and with object IDs obtained from verbal and/or gestural references only (Table 3). Also, the amount of signs used for each evaluation was varied by setting $Cmin$, the minimum number of communication segments in the training set in which a sign has occur to be considered. While testing the mixed approach, words were given a 50% penalty on their topic indication strength compared to object references. This was undertaken to enhance the results, since words were assumed to be less reliable in topic indicating than object references.

Table 1 shows the results for only word-based tracking. Although the results are not satisfying, three tendencies are observable: First,

Table 1: *Word based approach*

| CMin | 2 | 3 | 4 | 5 | 7 | 10 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|
| LSA | 64.7 | 64.2 | 64.0 | 64.2 | 64.3 | 62.3 | 61.1 | 57.2 |
| FSem | 58.3 | 60.4 | 59.8 | 61.6 | 63.0 | 60.5 | 59.1 | 53.8 |
| Vec. | 50.1 | 56.4 | 57.6 | 61.7 | 58.9 | 60.6 | 59.4 | 53.5 |

LSA performs better than the other approaches, with the simple vector space model as the worst. Second, the approach based on the simple vectorspace is somewhat susceptible to disturbances by rarely used words. With such words deleted, the performance of the model increases. Finally, the performance of each model drops when too many topic indicating cues are restricted, i.e. for high values of $Cmin$.

Table 2: *Mixed approach, penalty of 50% for words*

| CMin | 2 | 3 | 4 | 5 | 7 | 10 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|
| LSA | 80.6 | 81.5 | 82.6 | 81.2 | 81.0 | 80.3 | 73.6 | 63.8 |
| FSem | 69.4 | 72.8 | 69.9 | 66.5 | 71.1 | 69.9 | 73.8 | 64.6 |
| Vec. | 62.4 | 72.0 | 74.3 | 80.4 | 79.8 | 79.3 | 72.4 | 62.5 |

The same tendencies can be found in Table 2, where the simple vector space model sometimes performs even better than the Fuzzy Semantics approach. The fluctuation in the results of the Fuzzy Semantics Space are probably an indicator for the limited capability of the model to incorporate cues from modalities which share different distributions in the corpus. For topic tracking in HRI, it is desirable to cope with small values for $Cmin$, since more information is gained and smaller training sets are necessary.

Table 3: *References only*

| CMin | 2 | 3 | 4 | 5 | 7 | 10 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|
| LSA | 89.4 | 90.2 | 89.1 | 88.8 | 87.7 | 87.0 | 81.9 | 71.2 |
| FSem | 86.2 | 89.2 | 89.0 | 89.4 | 88.7 | 88.0 | 81.4 | 70.9 |
| Vec. | 90.0 | 90.0 | 88.9 | 90.0 | 87.4 | 86.2 | 82.1 | 71.3 |

Table 3 finally shows that topic tracking based only on multimodally resolved object references performs best on the corpus, indicating a strong connection of topics and objects in the environment. Interestingly, the three models show no great differences in performance, probably because of the fact that objects belong to topics nearly on a 1:1 basis, thus reducing "noise" tremendously.

## 5. Discussion and future work

The results show that using multimodal information is necessary for topic tracking on situated HRI. Obviously, object references are strong topic markers. The mixed model shows acceptable performance while delivering additional information compared to the reference based approach, namely word/topic relationships. Since this information is desirable, future work will comprise the development of mixed approaches, combining the good tracking results of solely object-based approaches with the additional information of mixed models. Solely word-based-tracking performs poorly, thus supporting our assumption.

Based on this offline system, an online topic tracking system for the robot BIRON has been developed. At the moment, experiments concerning topic tracking during communication sequences with BIRON are carried out. During this evaluation, the automatic segmentation will be tested. Future work will especially focus on the further development, integration and enhancement of the topic tracking software for the mobile robot.

## 6. Acknowledgements

The research has been supported by the European Union within the "Cognitive Robot Companion" (COGNIRON) project (FP6-IST-002020) and by the German Research Foundation within the Graduate Program "Task Oriented Communication".

## 7. References

[1] C.J. van Rijsbergen: "Information Retrieval". 2nd edition. URL: http://www.dcs.gla.ac.uk/Keith/Preface.html. 2005.

[2] J. Allan (ed.): "Topic Detection and Tracking". Kluwer Academic Publishers, Norwell, Massachusetts, 2002.

[3] Mikko Kurimo: "Thematic Indexing of Spoken Documents by Using Self-Organizing Maps". in: Speech Communication. Vol. 38. pp. 29-44.

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

[5] R. Serafin and B. Di Eugenio: "FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification". ACL, New York, 2004. pp. 692–699.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tür: "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics". Speech Communication, Vol. 1–2, 2000, pp. 127–154.

[7] G. Brown, G. Yule: "Discourse Analysis". Cambridge University Press, Cambridge, 1983.

[8] E. Leopold: "On Semiotic Spaces" in: Christian Wolff (ed.): LDV-Forum 18 (3). 2005.

[9] T.K. Landauer and S. T. Dumais: "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Aquisition, Induction and Representation of Knowledge", Psychological review, 1997, Vol. 104, pp. 211–240.

[10] G. Salton: "Automatic Information Organization and Retrieval". McGraw-Hill Book Company, New York, 1968.

[11] B. B. Rieger: "Connotative Dependency Structures in Semiotic Space" in: Rieger, B. (ed.): Empirical Semiotics II. Quantative Linguistics. Vol. 13, Bochum 1981, pp. 622-711.

[12] P. Cimiano, A. Hotho and S. Staab: "Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text" in: Proceedings of the European Conference on Artificial Intelligence (ECAI'04), 2004, pp 435–439.

[13] Jan F. Maas, B. Wrede: "BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction". Proceedings of the international conference on Language and Evaluation (LREC), Genoa, Italy. 2006. (to appear).

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

[15] H. Schmid: "Probabilistic Part-of-Speech Tagging Using Decision Trees"Proceedings of the International Conference on New Methods in Language Processing. 1994.