Crossmodal Cross-situational Learning with Attention*

Brigitte Krenn, Martin Trapp, Stephanie Gross, Friedrich Neubarth†

***Abstract*— We present experiments on a multimodal dataset of situated task descriptions annotated specifically for cross-modal object-word learning. In particular, we investigate effects of attentional cues incorporated in statistical learning models. Attentional cues help to direct listener's/learner's attention in multimodal communication, and thus facilitate learning the mapping between words in natural language and objects in the world. The results of our experiments indicate that, in the context of TAKE and PUT tasks, the object currently held by the instructor or is moved next to another object are important attentional cues. Further, we show that using the instructor's gaze as attention cue worsens the learning result for such tasks, which stands in contrast to previous work.**

I. INTRODUCTION

The overall goal of the presented work is to develop mechanisms enabling a robot to learn multimodal representations of previously unknown objects in situations of task-oriented communication. An important aspect of task-oriented communication is the temporal overlap of information conveyed across modalities, which in turn facilitates learning of word-referent relations. In this respect, task-related communication is comparable to parent - young infant communication. See for instance [6], [8] for the multisensory nature of early language learning. The scene descriptions which serve as input to the word-object labelling mechanisms consist of natural language utterances paired with lists of objects in attentional focus. For instance: There are several pieces of fruit on a plate. While the instructor utters *I take the strawberry*, she/he grasps and holds it. The visual focus is on the grasped object, the strawberry. We compare the effects of different visual attentional cues in combination with two variants of a simple learning algorithm based on conditional probabilities (baseline) and the model presented in [8] which integrates social cues and statistical learning.

The work is based on empirical data on multimodal task descriptions which are part of the (German) Multimodal Task Description (MMTD) Corpus [3]. For the learning experiments, we concentrate on Task 1, where 22 teachers explain to the camera how to (re-)order fruit (a banana, a strawberry and a pear). One after the other, the pieces of fruit are taken from a plate and put on certain locations on a white sheet of paper. The aim of this particular setting was to collect language and video data of simple actions, such as TAKE and PUT and combinations of these actions, to be used as input to crossmodal learning.

*The presented work is in part funded by WWTF project RALLI ICT15-045 and CHIST-ERA project ATLANTIS.

†The authors are with the Austrian Research Institute for Artificial Intelligence, Freyung 6, 1010 Vienna, Austria firstname.lastname@ofai.at

Videos and sound files are synchronised and manually annotated using Elan[1], including, amongst others, a transliteration of the utterance, eye gaze of the teacher, (left/right) hand touches object, object X moves next to object Y. It can be seen empirically that mentioning an object in speech typically co-occurs with an action that involves the object.

II. ATTENTION CUES

We qualitatively analysed the MMTD Corpus for attention directing cues. Especially Tasks 1, 3, 4 were used, as they are examples of humans explaining and conducting a task while another human or a robot (the learner) watches and listens. The full list of extracted attention cues comprises the following:

a) Verbal cues: 1. request for action, e.g. *du nimmst* ... ('you take ...'); 2. naming of objects, e.g. ... *du nimmst die Banane* ... ('... you take the banana ...'); 3. action and state verbs, e.g. *nehmen, legen, liegen* ('take, put, lie'); 4. spatial relations, locations and directions: PPs with spatial prepositions, space indexicals, e.g. ... *legst du die Banane daneben* ... ('you put the banana next to it'); 5. reference to perspective, e.g. *mit meiner/deiner linken Hand* ('with my/your left hand'); 6. particles, e.g. *ok, so, also, dazu, und zwar* ('ok, so, well, for that, namely'); 7. temporal adverbials draw attention to the timeline of an event, e.g. *zuerst, dann, jetzt* ('first, then, now').

b) Visual cues: 1. the object the teacher holds in her/his hands; 2. teacher gestures: pointing at, poising over, exhibiting an object; 3. moving object; 4. landing site of moved object; 5. posture shift of the speaker: lean forward = start explaining, lean backward = task finished. In general, changes in the visual field are indicators for what gets into the focus of visual attention.

As can be seen from a qualitative analysis of the data in Task 1, the major visual cues for directing the learner's attention are what the teacher holds in his/her hands and where, i.e. the landing site, the held object is moved to. This is in contrast to previous literature where pointing gestures and eye gaze are the prevalent cues included in computational models, e.g. [4], [5], [7]. Deictic gestures are rare in the data of Task 1 which focus on TAKE and PUT actions, hence we experiment with the following visual cues: objects which the teacher holds in her/his hands, objects next to which other objects are moved and eye gaze of the instructor.

III. WORD LEARNING EXPERIMENTS USING ATTENTION

c) Input Data: The 22 Task 1 description episodes are automatically segmented into subscenes based on a combina-

[1]https://tla.mpi.nl/tools/tla-tools/elan/

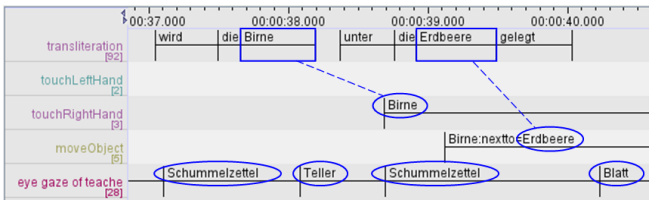

Fig. 1. Annotation example: annotation tiers and crossmodal relations

tion of manipulated object and the duration of speech pauses. Typically these utterances contain an action verb. From the resulting segments, pairs of 'transcribed utterance' and 'a set of objects in visual focus' are input to learning. In the experiments, we use four different sets of visual cues: VC1: only those objects the teacher takes into his/her hands in the current subscene; VC2: the touched object and object next to which the touched object is moved; VC3: eye gaze; VC4: touch, landing site and gaze are combined.

See Figure 1 for an annotation example. The instructor utters *... wird die Birne unter die Erdbeere gelegt ...* ('... the pear is put below the strawberry ...'). While the natural language input remains the same in all learning experiments, the visual input changes depending on which visual attention model is used: VC1 – {BIRNE}; VC2 – {BIRNE, ERDBEERE}; VC3 – {SCHUMMELZETTEL, TELLER, BLATT} (cheat sheet, plate, sheet of paper where the fruits are arranged on); VC4 – {BIRNE, ERDBEERE, SCHUMMELZETTEL, TELLER, BLATT}.

d) Learning Model: To achieve cross-situational object-word learning, we compare the attention-guiding cues VC1-4 using the unified learning model from [8] and frequentist representations $P(a|w)$ and $P(w|a)$ (baseline models). Whereby $w \in \mathcal{W}$ denotes a word in a scene and $a \in \mathcal{A} \subseteq \mathcal{O}$ denotes an object with attention on it. All three approaches are associative models for early word-learning. In comparison to raw conditional probabilities the model in [8] utilises additional latent alignment variables, similar to those in IBM Machine Translation Model 1 [1], and learns the association probabilities using an Expectation Maximisation (EM) algorithm. Therefore, the unified word learning model by [8] is expected to outperform a model exclusively based on the conditional probabilities $P(a|w)$. For comparison, we also computed $P(w|a)$.

e) Results: To assess the performance of different cross-situational object-word learning approaches, we computed the F_1-score for each combination of learning method and visual cues (VC). In this work we assume hard mappings between words and objects. The set of ground truth word-object mappings contains: *Erdbeere*-ERDBEERE, *Banane*-BANANE, *Birne*-BIRNE, *Teller*-TELLER, *Blatt*-BLATT. To estimate links between words and objects for each model, we take $\text{argmax}_w\, P(w|a)$ or $\text{argmax}_a\, P(a|w)$, respectively. The EM algorithm in [8] has been running until convergence for all experiments.

The results listed in Table I show F_1-scores computed for the above listed set of ground truth mappings. We found that

Attention	$P(a\|w)$	$P(w\|a)$	[8]
VC1 (touch hand)	0.4	**0.8**	0.4
VC2 (touch hand + next to)	0.6	0.5	**0.8**
VC3 (eye-gaze)	0.2	0.2	0.2
VC4 (all combined)	0.2	0.4	**0.8**

TABLE I

ESTIMATED F-1 SCORES FOR DIFFERENT ATTENTION CUES.

the unified word learning model by [8] performs best, except for the touch only condition. Furthermore, we observed that constructing mappings by using $P(w|a)$ performs superior to compute $P(a|w)$ in some cases, which matches the observations described in [2]. We also see that eye gaze is an unreliable cue for word-object mapping in situated task descriptions. Gaze is guided by varying information needs, e.g. looking at the cheat sheet (Task 1), backchanneling (Tasks 2, 4), planning or controlling the result of an action (Task 3).

IV. CONCLUSION AND ONGOING WORK

The presented crossmodal word learning experiments are the first stage of ongoing research for crossmodal word learning (objects and actions) based on multimodal input from task descriptions where a (human) instructor conducts and explains the task. Further experiments will be conducted with TAKE/PUT crossmodal data from other tasks and with additional actions. While the presented work operates on high-level data, i.e. transliterated speech and manually annotated visual information, in a further step, we will concentrate on learning word-referent mappings from low-level data. This data is coming from synchronized recordings of hand and object tracking, and speech. Suitable data sets are currently being collected.

REFERENCES

[1] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585, 2009.

[3] S. Gross and B. Krenn. The OFAI Multimodal Task Description Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1408–1414, 2016.

[4] C.-M. Huang and B. Mutlu. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 57–64. ACM, 2014.

[5] S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2):181–199, 2012.

[6] S. H. Suanda, L. B. Smith, and C. Yu. The multisensory nature of verbal discourse in parent–toddler interactions. *Developmental Neuropsychology*, 41(5-8):324–341, 2016.

[7] T. Williams, S. Acharya, S. Schreitter, and M. Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th International Conference on Human-Robot Interaction (HRI)*, pages 311–318, 2016.

[8] C. Yu and D. H. Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13):2149–2165, 2007.