Towards an Extrinsic Evaluation of Referring Expressions in Situated Dialogs

Philipp SPANGER IIDA Ryu TOKUNAGA Takenobu
{philipp,ryu-i,take}@cl.cs.titech.ac.jp

TERAI Asuka KURIYAMA Naoko
asuka@nm.hum.titech.ac.jp kuriyama@hum.titech.ac.jp

Tokyo Institute of Technology

Abstract

In the field of referring expression generation, while in the static domain both intrinsic and extrinsic evaluations have been considered, extrinsic evaluation in the dynamic domain, such as in a situated collaborative dialog, has not been discussed in depth. In a dynamic domain, a crucial problem is that referring expressions do not make sense without an appropriate preceding dialog context. It is unrealistic for an evaluation to simply show a human evaluator the whole period from the beginning of a dialog up to the time point at which a referring expression is used. Hence, to make evaluation feasible it is indispensable to determine an appropriate shorter context. In order to investigate the context necessary to understand a referring expression in a situated collaborative dialog, we carried out an experiment with 33 evaluators and a Japanese referring expression corpus. The results contribute to finding the proper contexts for extrinsic evalution in dynamic domains.

1 Introduction

In recent years, the NLG community has paid significant attention to the task of generating referring expressions, reflected in the seting-up of several competitive events such as the TUNA and GIVE-Challenges at ENLG 2009 (Gatt et al., 2009; Byron et al., 2009).

With the development of increasingly complex generation systems, there has been heightened interest in and an ongoing significant discussion on different evaluation measures for referring expressions. This discussion is carried out broadly in the field of generation, including in the multi-modal domain, e.g. (Stent et al., 2005; Foster, 2008). Figure 1 shows a schematic overview of recent work on evaluation of referring expressions along the two axes of evaluation method and domain in which referring expressions are used.

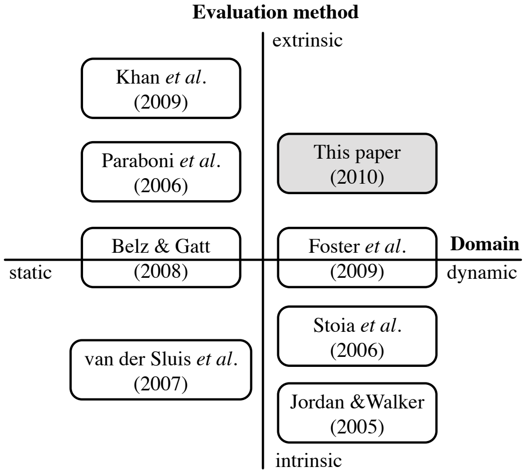


Figure 1: Overview of recent work on evaluation of referring expressions

There are two different evaluation methods corresponding to the bottom and the top of the vertical axis in Figure 1: *intrinsic* and *extrinsic* evaluations (Sparck Jones and Galliers, 1996). Intrinsic methods often measure similarity between the system output and the gold standard corpora using metrics such as tree similarity, string-edit-distance and BLEU (Papineni et al., 2002). Intrinsic methods have recently become popular in the NLG community. In contrast, extrinsic methods evaluate generated expressions based on an external metric, such as its impact on human task performance.

While intrinsic evaluations have been widely used in NLG, e.g. (Reiter et al., 2005), (Cahill and van Genabith, 2006) and the competitive 2009 TUNA-Challenge, there have been a number of criticisms against this type of evaluation. (Reiter and Sripada, 2002) argue, for example, that generated text might be very different from a corpus but still achieve the specific communicative goal.

An additional problem is that corpus-similarity metrics measure how well a system reproduces what speakers (or writers) do, while for most NLG systems ultimately the most important consideration is its effect on the human user (i.e. listener or reader). Thus (Khan et al., 2009) argues that "measuring *human-likeness* disregards effectiveness of these expressions".

Furthermore, as (Belz and Gatt, 2008) state "there are no significant correlations between intrinsic and extrinsic evaluation measures", concluding that "similarity to human-produced reference texts is not necessarily indicative of quality as measured by human task performance".

From early on in the NLG community, task-based extrinsic evaluations have been considered as the most meaningful evaluation, especially when having to convince people in other communities of the usefulness of a system (Reiter and Belz, 2009). Task performance evaluation is recognized as the "only known way to measure the effectiveness of NLG systems with real users" (Reiter et al., 2003). Following this direction, the GIVE-Challenges (Koller et al., 2009) at INLG 2010 (instruction generation) also include a task-performance evaluation.

In contrast to the vertical axis of Figure 1, there is the horizontal axis of the domain in which referring expressions are used. Referring expressions can thus be distinguished according to whether they are used in a *static* or a *dynamic* domain, corresponding to the left and right of the horizontal axis of Figure 1. A static domain is one such as the TUNA corpus (van Deemter, 2007), which collects referring expressions based on a motionless image. In contrast, a dynamic domain comprises a constantly changing situation where humans need context information to identify the referent of a referring expression.

Referring expressions in the static domain have been evaluated relatively extensively. A recent example of an intrinsic evaluation is (van der Sluis et al., 2007), who employed the Dice-coefficient measuring corpus-similarity. There have been a number of extrinsic evaluations as well, such as (Paraboni et al., 2006) and (Khan et al., 2009), respectively measuring the effect of overspecification on task performance and the impact of generated text on accuracy as well as processing speed. They belong thus in the top-left quadrant of Figure 1.

Over a recent period, research in the generation of referring expressions has moved to dynamic domains such as situated dialog, e.g. (Jordan and Walker, 2005) and (Stoia et al., 2006). However, both of them carried out an intrinsic evaluation measuring corpus-similarity or asking evaluators to compare system output to expressions used by human (the right bottom quadrant in Figure 1).

The construction of effective generation systems in the dynamic domain requires the implementation of an extrinsic task performance evaluation. There has been work on extrinsic evaluation of instructions in the dynamic domain on the GIVE-2 challenge (Byron et al., 2009), which is a task to generate instructions in a virtual world. It is based on the GIVE-corpus (Gargett et al., 2010), which is collected through keyboard interaction. The evaluation measures used are e.g. the number of successfully completed trials, completion time as well as the numbers of instructions the system sent to the user. As part of the JAST project, a Joint Construction Task (JCT) puzzle construction corpus (Foster et al., 2008) was created which is similar in some ways in its set-up to the REX-J corpus which we use in the current research. There has been some work on evaluating generation strategies of instructions for a collaborative construction task on this corpus, both considering intrinsic as well as extrinsic measures (Foster et al., 2009). Their main concern is, however, the interaction between the text structure and usage of referring expressions. Therefore, their "context" was given a priori.

However, as can be seen from Figure 1, in the field of referring expression generation, while in the static domain both intrinsic and extrinsic evaluations have been considered, the question of realizing an extrinsic evaluation in the dynamic domain has not been dealt with in depth by previous work. This paper addresses this shortcoming of previous work and contributes to "filling in" the missing quadrant of Figure 1 (the top-right).

The realization of such an extrinsic evaluation faces one key difficulty. In a static domain, an extrinsic evaluation can be realized relatively easily by showing evaluators the *static* context (e.g. any image) and a referring expression, even though this is still costly in comparison to intrinsic methods (Belz and Gatt, 2008).

In contrast, an extrinsic evaluation in the *dynamic* domain needs to present an evaluator with the *dynamic* context (e.g. a certain length of the recorded dialog) preceding a referring expression. It is clearly not feasible to simply show the *whole* preceding dialog; this would make even a very small-scale evaluation much too costly. Thus, in order to realize a cost-effective extrinsic evaluation in a dynamic domain we have to deal with the additional parameter of time length and content of the context shown to evaluators.

This paper investigates the context necessary for humans to understand different types of referring expressions in a situated domain. This work thus charts new territory and contributes to developing a extrinsic evaluation in a dynamic domain. Significantly, we consider not only linguistic but also extra-linguistic information as part of the context, such as the actions that have been carried out in the preceding interaction. Our results indicate that, at least in this domain, extrinsic evaluation results in dynamic domains can depend on the specific amount of context shown to the evaluator. Based on the results from our evaluation experiments, we discuss the broader conclusions to be drawn and directions for future work.

2 Referring Expressions in the REX-J Corpus

We utilize the REX-J corpus, a Japanese corpus of referring expressions in a situated collaborative task (Spanger et al., 2009a). It was collected by recording the interaction of a pair of dialog participants solving the Tangram puzzle cooperatively. The goal of the Tangram puzzle is to construct a given shape by arranging seven pieces of simple figures as shown in Figure 2

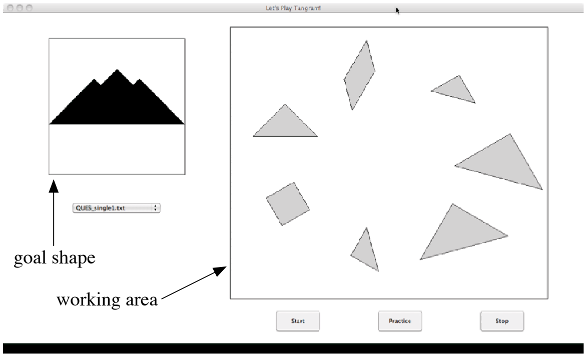


Figure 2: Screenshot of the Tangram simulator

In order to record the precise position of every piece and every action by the participants, we implemented a simulator. The simulator displays two areas: a goal shape area, and a working area where pieces are shown and can be manipulated.

We assigned different roles to the two participants of a pair: *solver* and *operator*. The solver can see the goal shape but cannot manipulate the pieces and hence gives instructions to the operator; by contrast, the operator can manipulate the pieces but can not see the goal shape. The two participants collaboratively solve the puzzle sharing the working area in Figure 2.

In contrast to other recent corpora of referring expressions in situated collaborative tasks (e.g. COCONUT corpus (Di Eugenio et al., 2000) and SCARE corpora (Byron et al., 2005)), in the REX-J corpus we allowed comparatively large real-world flexibility in the actions necessary to achieve the task (such as flipping, turning and moving of puzzle pieces at different degrees), relative to the task complexity. The REX-J corpus thus allows us to investigate the interaction of linguistic and extra-linguistic information. Interestingly, the GIVE-2 challenge at INLG 2010 notes its “main novelty” is allowing “continuous moves rather than discrete steps as in GIVE-1”. Our work is in line with the broader research trend in the NLG community of trying to get away from simple “discrete” worlds to more realistic settings.

The REX-J corpus contains a total of 1,444 tokens of referring expressions in 24 dialogs with a total time of about 4 hours and 17 minutes. The average length of each dialog is 10 minutes 43 seconds. The asymmetric data-collection setting encouraged referring expressions from the solver (solver: 1,244 tokens, operator: 200 tokens). We exclude from consideration 203 expressions referring to either groups of pieces or whose referent cannot be determined due to ambiguity, thus leaving us 1,241 expressions.

We identified syntactic/semantic features in the collected referring expressions as listed in Table 1: (a) demonstratives (adjectives and pronouns), (b) object attribute-values, (c) spatial relations and (d) actions on an object. The underlined part of the examples denotes the feature in question.

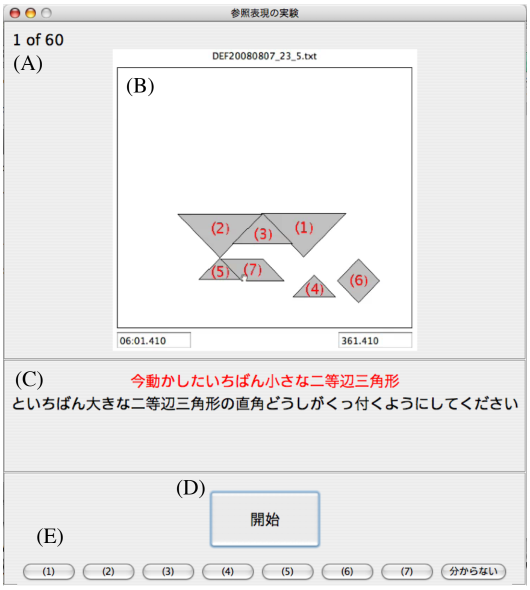
3 Design of Evaluation Experiment

The aim of our experiment is to investigate the “context” (content of the time span of the recorded

Table 1: Syntactic and semantic features of referring expressions in the REX-J corpus

Feature	Tokens	Example
(a) demonstrative	742	*ano migigawa no sankakkei* (that triangle at the right side)
(b) attribute	795	*tittyai sankakkei* (the small triangle)
(c) spatial relations	147	*hidari no okkii sankakkei* (the small triangle on the left)
(d) action-mentioning	85	*migi ue ni doketa sankakkei* (the triangle you put away to the top right)

interaction prior to the uttering of the referring expression) necessary to enable successful identification of the referent of a referring expression. Our method is to vary the context presented to evaluators and then to study the impact on human referent identification. In order to realize this, for each instance of a referring expression, we vary the length of the video shown to the evaluator.



今動かしたいちばん小さな二等辺三角形といちばん大きな二等辺三角形の直角どうしがくっ付くようにしてください

(A): Counter (1-60)
(B): Video of shared working area in the simulator
(C): Utterance including the referring expression to evaluate (shown in red)
(D): Start/repeat button
(E): Selection buttons (1-7) and "I don't know"-button

Figure 3: The interface presented to evaluators

The basic procedure of our evaluation experiment is as follows:

(1) present human evaluators with speech and video from a dialog that captures shared working area of a certain length previous to the uttering of a referring expression,
(2) stop the video and display as text the next solver's utterance including the referring expression (shown in red),
(3) ask the evaluator to identify the referent of the presented referring expression (if the evaluator wishes, he/she can replay the video as many times as he likes),
(4) proceed to the next referring expression (go to (1)).

Figure 3 shows a screenshot of the interface prepared for this experiment.

The test data consists of three types of referring expressions: DPs (demonstrative pronouns), AMEs (action-mentioning expressions), and OTHERs (any other expression that is neither a DP nor AME, e.g intrinsic attributes and spatial relations). DPs are the most frequent type of referring expression in the corpus. AMEs are expressions that utilize an action on the referent such as "the triangle you put away to the top right" (see Table 1)[1]. As we pointed out in our previous paper (Spanger et al., 2009a), they are also a fundamental type of referring expression in this domain.

The basic question in investigating a suitable context is what information to consider about the preceding interaction; i.e. over what parameters to vary the context. In previous work on the generation of demonstrative pronouns in a situated domain (Spanger et al., 2009b), we investigated the role of linguistic and extra-linguistic information, and found that time distance from the last action (LA) on the referent as well as the last mention (LM) to the referent had a significant influence on the usage of referring expressions. Based on those results, we focus on the information on the referent, namely LA and LM.

For both AMEs and OTHERs, we only consider two possibilities of the order in which LM and LA appear before a referring expression (REX), depending on which comes first. These are shown in Figure 4, context patterns (a) LA-LM and (b) LM-LA. Towards the very beginning of a dialog, some referring expressions have no LM and LA; those expressions are not considered in this research.

All instances of AMEs and OTHERs in our test data belong to either the LA-LM or the LM-LA

[1] An action on the referent is usually described by a verb as in this example. However, there are cases with a verb ellipsis. While this would be difficult in English, it is natural and grammatical in Japanese.

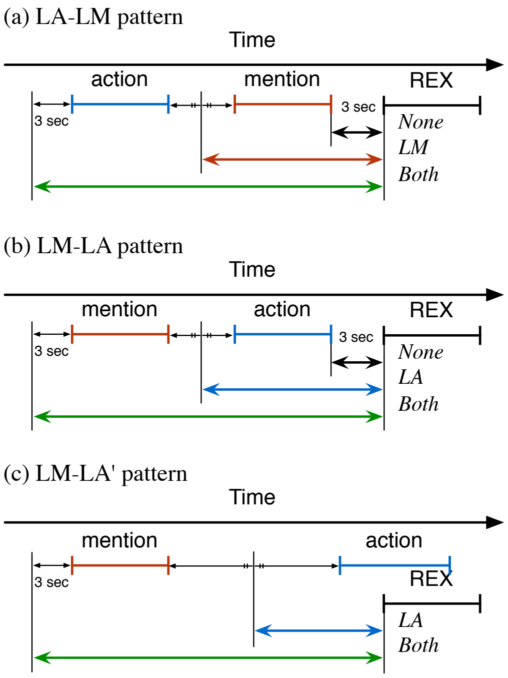


Figure 4: Schematic overview of the three context Patterns

pattern. For each of these two context patterns, there are three possible contexts[2]: *Both* (including both LA and LM), *LA/LM* (including either LA or LM) and *None* (including neither). Depending on the order of LA and LM prior to an expression, only one of the variations of *LA/LM* is possible (see Figure 4 (a) and (b)).

In contrast, DPs tend to be utilized in a deictic way in such situated dialogs (Piwek, 2007). We further noted in (Spanger et al., 2009b), that DPs in a collaborative task are also frequently used when the referent is under operation. While they belong neither to the LA-LM nor the LM-LA pattern, it would be inappropriate to exclude those cases. Hence, for DPs we consider another situation where the last action on the referent *overlaps* with the utterance of the DP (Figure 4 (c) LM-LA' pattern). In this case, we consider an ongoing operation on the referent as a "last action". Another peculiarity of the LM-LA' pattern is that we have no *None* context in this case, since there is no way to show a video without showing LA (the current operation).

Given the three basic variations of context, we recruited 33 university students as evaluators and divided them equally into three groups, i.e. 11 evaluators per group. As for the referring expressions to evaluate, we selected 60 referring expressions used by the solver from the REX-J corpus (20 from each category), ensuring all were correctly understood by the operator during the recorded dialog. We selected those 60 instances from expressions where both LM and LA appeared within the last 30 secs previous to the referring expression. This selection excludes initial mentions, as well as expressions where only LA or only LM exists or they do not appear within 30 secs. Hence the data utilized for this experiment is limited in this sense. We need further experiments to investigate the relation between the time length of contexts and the accuarcy of evaluators. We will return to this issue in the conclusion.

We combined 60 referring expressions and the three contexts to make the test instances. Following the Latin square design, we divided these test instances into three groups, distributing each of the three contexts for every referring expression to each group. The number of contexts was uniformly distributed over the groups. Each instance group was assigned to each evaluator group.

For each referring expression instance, we record whether the evaluator was able to correctly identify the referent, how long it took them to identify it and whether they repeated the video (and if so how many times).

Reflecting the distribution of the data available in our corpus, the number of instances per context pattern differs for each type of referring expression. For AMEs, overwhelmingly the last action on the referent was more recent than the last mention. Hence we have only two LA-LM patterns among the 20 AMEs in our data. For OTHERs, the balance is 8 to 12, with a slight majority of LM-LA patterns. For DPs, there is a strong tendency to use a DP when a piece is under operation (Spanger et al., 2009b). Of the 20 DPs in the data, 2 were LA-LM, 5 were LM-LA pattern while 13 were of the LM-LA' pattern (i.e. their referents were under operation at the time of the utterance). For these 13 instances of LM-LA' we do not have a *None* context.

The average stimulus times, i.e. time period of presented context, were 7.48 secs for *None*, 11.04 secs for *LM/LA* and 18.10 secs for *Both*.

[2] To be more precise, we set a margin at the beginning of contexts as shown in Figure 4.

Table 2: Accuracy of referring expression identification per type and context

Type	context pattern\Context	*None*	*LM/LA*	*Both*	Increase [*None* → *Both*]
DP	(LA-LM)	0.909 (20/22)	0.955 (21/22)	0.955 (21/22)	0.046
	(LM-LA)	0.455 (25/55)	0.783 (155/198)	0.843 (167/198)	0.388
Total		0.584	0.800	0.855	0.271
AME	(LA-LM)	0.227 (5/22)	0.455 (10/22)	0.682 (15/22)	0.455
	(LM-LA)	0.530 (105/198)	0.859 (170/198)	0.879 (174/198)	0.349
Total		0.500	0.818	0.859	0.359
OTHER	(LA-LM)	0.784 (69/88)	0.852 (75/88)	0.943 (83/88)	0.159
	(LM-LA)	0.765 (101/132)	0.788 (104/132)	0.879 (116/132)	0.114
Total		0.773	0.814	0.905	0.132
Overall		0.629 (325/517)	0.811 (535/660)	0.903 (576/638)	0.274

4 Results and Analysis

In this section we discuss the results of our evaluation experiment. In total 33 evaluators participated in our experiment, each solving 60 problems of referent identification. Taking into account the absence of the *None* context for the DPs of the LM-LA' pattern (see (c) in Figure 4), we have 1,815 responses to analyze. We focus on the impact of the three contexts on the three types of referring expressions, considering the two context patterns LA-MA and LM-LA.

4.1 Overview of Results

Table 2 shows the micro averages of the accuracies of referent identification of all evaluators over different types of referring expressions with different contexts. Accuracies increase with an increase in the amount of information in the context; from *None* to *Both* by between 13.2% (OTHERs) and 35.9% (AMEs). The average increase of accuracy is 27.4%.

Overall, for AMEs the impact of the context is the greatest, while for OTHERs it is the smallest. This is not surprising given that OTHERs tend to include intrinsic attributes of the piece and its spatial relations, which are independent of the preceding context. We conducted ANOVA with the context as the independent variable, testing its effect on identification accuracy. The main effect of the context was significant on accuracy with $F(2, 1320) = 9.17$, $p < 0.01$. Given that for DPs we did not have an even distribution between contexts, we only utilized the results of AMEs and OTHERs.

There are differences between expression types in terms of the impact of addition of LM/LA into the context, which underlines that when studying context, the relative role and contribution of LA and LM (and their interaction) must be looked at in detail for different types of referring expressions.

Over all referring expressions, the addition into a *None* context of LM yields an average increase in accuracy of 9.1% for all referring expression types, while for the same conditions the addition of LA yields an average increase of 21.3%. Hence, interestingly for our test data, the addition of LA to the context has a positive impact on accuracy by more than two times over the addition of LM.

It is also notable that even with neither LA nor LM present (i.e. the *None* context), the evaluators were still able to correctly identify referents in between 50–68.6% (average: 62.9%) of the cases. While this accuracy would be insufficient for the evaluation of machine generated referring expressions, it is still higher than one might expect and further investigation of this case is necessary.

4.2 Demonstrative Pronouns

For DPs, there is a very clear difference between the two patterns (LM-LA and LA-LM) in terms of the increase of accuracy with a change of context. While accuracy for the LA-LM pattern remains at a high level (over 90%) for all three contexts (and there is only a very small increase from *None* to *Both*), for the LM-LA pattern there is a strong increase from *None* to *Both* of 38.8%.

The difference in accuracy between the two context patterns of DPs in the *None* context might come from the mouse cursor effect. The two expressions of LA-LM pattern happened to have a mouse cursor on the referent, when they were used, resulting in high accuracy. On the other hand, 4 out of 5 expressions of LM-LA pattern did not have a mouse cursor on the referent. We have currently no explanation for the relation between context patterns and the mouse position. While we have only 7 expressions in the *None* context for DPs and hence cannot draw any decisive conclusions, we note that the impact of the mouse position is a likely factor.

For the LM-LA pattern, there is an increase in accuracy of 32.8% from *None* to the *LA*-context. Overwhelmingly, this represents instances in which the referents are being operated at the point in time when the solver utters a DP (this is in fact the LM-LA' pattern, which has no *None* context). For those instances, the current operation information is sufficient to identify the referents. In contrast, addition of LM leads only to a small increase in accuracy of 5.6%. This result is in accordance with our previous work on the generation of DPs, which stressed the importance of extra-linguistic information in the framework of considering the interaction between linguistic and extra-linguistic information.

4.3 Action-mentioning Expressions

While for AMEs the number of instances is very uneven between patterns (similar to the distribution for DPs), there is a strong increase in accuracy from the *None* context to the *Both* context for both patterns (between 30% to almost 50%). However, there is a difference between the two patterns in terms of the relative contribution of LM and LA to this increase.

While for the LA-LM pattern the impact of adding LM and LA is very similar, for the LM-LA pattern the major increase in accuracy is due to adding LA into the *None* context. This indicates that for AMEs, LA has a stronger impact on accuracy than LM, as is to be expected. The strong increase for AMEs of the LM-LA pattern when adding LA into the context is not surprising, given that the evaluators were able to see the action mentioned in the AME.

For the opposite reason, it is not surprising that AMEs show the lowest accuracy in the *None* context, given that the last action on the referent is not seen by the evaluators. However, accuracy was still slightly over 50% in the LM-LA pattern. Overall, of the 18 instances of AMEs of the LM-LA pattern, in the *None* context a majority of evaluators correctly identified 9 and erred on the other 9. Further analysis of the difference between correctly and incorrectly identified AMEs led us to note again the important role of the mouse cursor also for AMEs.

Comparing to the LM-LA pattern, we had very low accuracy even with the Both context. As we mentioned in the previous section, we had very skewed test instances for AME, i.e. 18 LM-LA patterns vs. 2 LA-LM patterns. We need further investigation on the LA-LM pattern of AME with more large number of instances.

Of the 18 LM-LA instances of AMEs, there are 14 instances that mention a verb describing an action on the referent. The referents of 6 of those 14 AMEs were correctly determined by the evaluators and in all cases the mouse cursor played an important role in enabling the evaluator to determine the referent. The evaluators seem to utilize the mouse position at the time of the uttering of the referring expression as well as mouse movements in the video shown. In contrast, for 8 out of the 9 incorrectly determined AMEs no such information from the mouse was available. There was a very similar pattern for AMEs that did not include a verb. These points indicate that movements and the position of the mouse both during the video as well as the time point of the uttering of the referring expression give important clues to evaluators.

4.4 Other Expressions

There is a relatively even gain in identification accuracy from *None* to *Both* of between about 10–15% for both patterns. However, there is a similar tendency as for AMEs, since there is a difference between the two patterns in terms of the relative contribution of LM and LA to this increase. While for the LA-LM pattern the impact of adding LM and LA is roughly equivalent, for the LM-LA pattern the major increase in accuracy is due to adding LM into the LA-context.

For this pattern of OTHERs, LM has a stronger impact on accuracy than LA, which is exactly the opposite tendency to AMEs. For OTHERs (e.g. use of attributes for object identification), seeing the last action on the target has a less positive impact than listening to the last linguistic mention.

Furthermore, we note the relatively high accuracy in the *None* context for OTHERs, underlining the context-independence of expressions utilizing attributes and spatial relations of the pieces.

4.5 Error Analysis

We analyzed those instances whose referents were not correctly identified by a majority of evaluators in the *Both* context. Among the three expression types, there were about 13–16% of wrong answers. In total for 7 of the 60 expressions a majority of evaluators gave wrong answers (4 DPs, 2 AMEs and 1 OTHER). Analysis of these instances indicates that some improvements of our conception of "context" is needed.

For 3 out of the 4 DPs, the mouse was not over the referent or was closer to another piece. In addition, these DPs included expressions that pointed to the role of a piece in the overall construction of the goal shape, e.g. "*soitu ga atama* (that is the head)", or where a DP is used as part of a more complex referring expression, e.g. "*sore to onazi katati* … (the same shape as this)", intended to identify a different piece. For a non-participant of the task, such expressions might be difficult to understand in any context. This phenomenon is related to the "overhearer-effect" (Schober et al., 1989).

The two AMEs that the majority of evaluators failed to identify in the *Both* context were also misidentified in the *LA* context. Both AMEs were missing a verb describing an action on the referent. While for AMEs including a verb the accuracy increased from *None* to *Both* by 50%, for AMEs without a verb there was an increase by slightly over 30%, indicating that in the case where an AME lacks a verb, the context has a smaller positive impact on accuracy than for AMEs that include a verb. In order to account for those cases, further work is necessary, such as investigating how to account for the information on the distractors.

5 Conclusions and Future Work

In order to address the task of designing a flexible experiment set-up with relatively low cost for extrinsic evaluations of referring expressions, we investigated the context that needs to be shown to evaluators in order to correctly determine the referent of an expression.

The analysis of our results showed that the context had a significant impact on referent identification. The impact was strongest for AMEs and DPs and less so for OTHERs. Interestingly, we found for both DPs and AMEs that including LA in the context had a stronger positive impact than including LM. This emphasizes the importance of taking into account extra-linguistic information in a situated domain, as considered in this study.

Our analysis of those expressions whose referent was incorrectly identified in the *Both* context indicated some directions for improving the "context" used in our experiments, for example looking further into AMEs without a verb describing an action on the referent. Generally, there is a necessity to account for mouse movements during the video shown to evaluators as well as the problem for extrinsic evaluations of how to address the "overhearer's effect".

While likely differing in the specifics of the set-up, the methodology in the experiment design discussed in this paper is applicable to other domains, in that it allows a low-cost flexible design of evaluating referring expressions in a dynamic domain. In order to avoid the additional effort of analyzing cases in relation to LM and LA, in the future it will be desirable to simply set a certain time period and base an evaluation on such a set-up.

However, we cannot simply assume that a longer context would yield a higher identification accuracy, given that evaluators in our set-up are not actively participating in the interaction. Thus there is a possibility that identification accuracy actually decreases with longer video segments, due to a loss of the evaluator's concentration. Further investigation of this question is indicated.

Based on the work reported in this paper, we plan to implement an extrinsic task-performance evaluation in the dynamic domain. Even with the large potential cost-savings based on the results reported in this paper, extrinsic evaluations will remain costly. Thus one important future task for extrinsic evaluations will be to investigate the correlation between extrinsic and intrinsic evaluation metrics. This in turn will enable the use of cost-effective intrinsic evaluations whose results are strongly correlated to task-performance evaluations. This paper made an important contribution by pointing the direction for further research in extrinsic evaluations in the dynamic domain.

References

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200.

Donna Byron, Thomas Mampilly, Vinay Sharma, and Tianfang Xu. 2005. Utilizing visual attention for cross-modal coreference interpretation. In *CONTEXT 2005*, pages 83–96.

Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 165–173.

Aoife Cahill and Josef van Genabith. 2006. Robust PCFG-based generation using automatically acquired lfg approximations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1033–1040.

Barbara Di Eugenio, Pamela. W. Jordan, Richmond H. Thomason, and Johanna. D Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6):1017–1076.

Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L. Hill, Jon Oberlander, and Alois Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of 3rd Human-Robot Interaction*, pages 295–302.

Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI 2009)*, pages 1818–1823.

Mary Ellen Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proceedings of the 5th International Natural Language Generation Conference (INLG 2008)*, pages 95–103.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The give-2 corpus of giving instructions in virtual environments. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 2401–2406.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 174–182.

Pamela W. Jordan and Marilyn A. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Imtiaz Hussain Khan, Kees van Deemter, Graeme Ritchie, Albert Gatt, and Alexandra A. Cleland. 2009. A hearer-oriented evaluation of referring expression generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 98–101.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Sara Dalzel-Job, Jon Oberlander, and Johanna Moore. 2009. Validating the web-based evaluation of nlg systems. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 301–304.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Ivandré Paraboni, Judith Masthoff, and Kees van Deemter. 2006. Overspecified reference in hierarchical domains: Measuring the benefits for readers. In *Proceedings of the 4th International Natural Language Generation Conference (INLG 2006)*, pages 55–62.

Paul L.A. Piwek. 2007. Modality choise for generation of referring acts. In *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*, pages 129–139.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceesings of 2nd International Natural Language Generation Conference (INLG 2002)*, pages 97–104.

Ehud Reiter, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a failure: generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

Michael F. Schober, Herbert, and H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. 2009a. A Japanese corpus of referring expressions used in a situated collaboration task. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 110 – 113.

Philipp Spanger, Masaaki Yasuhara, Iida Ryu, and Tokunaga Takenobu. 2009b. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Linguistics and Intelligent Text Processing*, pages 341–351. Springer-Verlag.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Natural Language Generation Conference (INLG 2006)*, pages 81–88.

Kees van Deemter. 2007. TUNA: Towards a unified algorithm for the generation of referring expressions. Technical report, Aberdeen University. www.csd.abdn.ac.uk/research/tuna/pubs/TUNA-final-report.pdf.

Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of Recent Advances in Natural Languae Processing (RANLP 2007)*.