Measuring Relative Target User Group Success in Spoken Conversation for Edutainment

Niels Ole Bernsen
Natural Interactive Systems Lab
Campusvej 55, DK 5230 Odense M Denmark
+45 6550 3544, nob@nis.sdu.dk

ABSTRACT

The paper presents corpus data obtained from a relatively large field test of a Wizard of Oz (WoZ)-simulated specification of a multimodal domain-oriented spoken conversation system for edutainment. As the system design targets 10-18 years old users, a metrics is proposed for measuring the extent to which the simulated system specifically manages to appeal to its target user group. The metrics are applied to the WoZ corpus data, focusing on how to handle the observed differences between native and non-native English speaking users. This leads to a derived metrics which seems useful for system development progress evaluation.

Keywords

Evaluation metrics for edutainment systems, animated agent systems evaluation, multimodal spoken dialogue systems.

INTRODUCTION

This paper presents corpus-based results on the extent to which we have reached our target user group in a system aimed to have edutaining conversation with primarily 10-18 years old users. The system enables spoken English domain oriented conversation between users and life-like embodied fairytale author Hans Christian Andersen (HCA) and is being developed in the EU NICE project on Natural Interactive Communication for Edutainment [2]. Based on the design specification of the first system prototype, a Wizard of Oz (WoZ) simulation was carried out in the summer of 2003 at the HCA Museum in his native city, Odense, Denmark. During 10 days, approx. 500 conversations were recorded yielding 30 hours of spoken conversation data. This data has been transcribed and transcription coded. Each conversation has been evaluated with respect to the English language proficiency of the user. Topic-tagging of the corpus is in progress in order to identify all conversational topics addressed in the corpus.

By contrast with task-oriented spoken dialogue systems, whether unimodal (speech-only) or multimodal, domain-oriented systems do not help the user accomplish any particular task(s). Rather, the user can talk to the system spontaneously about anything, in any order, within the system's knowledge domain(s). Such systems raise novel issues of corpus-based evaluation, in particular, perhaps, if they have entertainment as one of their primary goals. For instance, classical dialogue efficiency metrics are probably irrelevant to their evaluation [1,3]. Rather, issues such as entertainment and edutainment success move to the forefront.

In the NICE HCA system, the target user group is 10-18 years old kids and adolescents. It is important to be able to evaluate the extent to which the target users are actually being entertained by the system, both in absolute terms and relative to non-target system users. This paper addresses the latter, relative, evaluation issue.

In the following, we briefly describe the NICE HCA system specification and the WoZ simulation. We then propose a metrics for measuring target group success in conversational systems for edutainment, discuss how to apply the metrics when the large majority of the users are non-native English speakers, present the results of applying the metrics, and discuss how to use the metrics for progress evaluation during continued system development.

NICE HCA SYSTEM SPECIFICATION

The system specification which was WoZ-simulated provides HCA with six domains of conversation: the childhood part of his life, the fairytale part of his work, his personality and visible physical presence in his study, gathering knowledge about the user, and his role as "gatekeeper" for access to the fairytale world in which users can interact with some of his fairytale characters. In addition, HCA has the "meta" domain of handling meta-communication caused by, e.g., user repeat requests or low input confidence scores. The following system aspects were not simulated: (i) the details of the system's error-handling meta-communication had not been specified at the time. In general, realistic system error behaviour as well as user and system error handling behaviour tend to be difficult to simulate using WoZ [1]; (ii) due to limitations of the graphical animation platform at the time, it was not possible to simulate the 2D user gesture input and its processing which form part of the (now implemented and running) first NICE HCA prototype; (iii) for the same reason, the simulation did not include HCA's conversational listening behaviour which, in the first prototype, enables HCA to show real-time attention to the user's spoken and gesture input. Finally, (iv) the details on exactly when HCA would exhibit emotional behaviour had not been designed at the time.

WIZARD OF OZ SIMULATION DETAILS

In technical Wizard of Oz terms, the simulation may be described as a full, field, close-to-complete specification, messy-experiment WoZ. A *full WoZ simulation* is one which does not include any implemented system components. A *field WoZ simulation* is conducted in the field rather than in a controlled laboratory setting. Users simply walk up and use the system with little or no introduction to its purpose or capabilities, and no requirements on the users whatsoever. A *close-to-complete specification WoZ simulation* is based on the system specification rather than on a more or less loosely defined purpose of gathering interesting data for a system which still has to be specified or which is under specification. The non-simulated specification aspects were described above. Finally, a *messy-experiment WoZ simulation* is one in which interaction experimentation is being carried out under less than strict textbook experimental conditions. Thus, in the simulations reported, the wizards were instructed to make, at their discretion, particular kinds of conversational improvisations which went beyond the system specification. These improvisations served as "messy" experiments intended to elicit user behaviours in addition to those which would be elicited by an uncompromising adherence to the system specification. For example, the wizards could talk out-of-specification in order to query the users about technical inventions made after HCA's times.

Figure 1. HCA addressing the user.

In the Museum, HCA was installed on a laptop which was wirelessly connected to the wizard working in the basement. A student had the task to round up kids and adolescents, inviting them to talk to "a nice person". In addition, a small poster in Danish and English invited the 10-18 year olds and other visitors to talk to this person, describing the system as a spoken computer game and informing users that their conversations would be recorded for research purposes. The user just had to don the headset and get started. Two wizards took turns simulating HCA through speech and movement control. Their main support was a hypertext document organised hierarchically by domain and topic, enabling quick navigation to find appropriate output to the user in the discourse context. The wizards were trained in advance, the training being supported by a written Wizard's guide, and instructed to make notes which were discussed in day-to-day briefing sessions.

Figure 2. A wizard in action.

BASIC DATA

The basic turn-level simulation data are shown in Table 1. Turn numbers measure the total number of turns made by the user and HCA in a conversation. Since they take turns communicating, each of them will produce half of the turns +/- a single turn.

The total of 498 conversations excludes four conversations of <4 turns and two conversations in which the transcribers thoroughly mixed up the users. All other recorded conversations are included in Table 1. The reason for leaving out the <4 turns exchanges is that there is hardly any conversation if what happens is merely a user saying, e.g., "Hello" and HCA responding, e.g., "Hello, welcome to my study". The reason why Table 1 provides information on users' age, gender and nationality, is that HCA has as a priority in conversation to gather this information from the users in order to use it as the conversation proceeds. He will thus try to collect this information either up front or, at least, early on in each conversation. Age information was provided by 91.0% of all users, gender information by 89.2%, and nationality information by 87.1%. The most common reason, by far, for not providing age, gender, and/or nationality information was that the user broke off the conversation before HCA could gather this data. This is evidenced by the facts that the average number of turns for age-unknown users is as low as 13 and the average number of turns of gender-unknown users is similarly low at 14 (Table 1). In a few cases, the wizards forgot to ask for the information. Few users refused to tell HCA their age or gender, and only in a couple of cases is there reason to believe that a user gave deliberately wrong information. An example is Maria on Day 9 who first had a 98-turn conversation as Maria, an 11 years old female from Denmark, and then came back to have a 24-turn conversation as Maria, a 13 years old boy from Denmark wanting to discuss girls with HCA, unfortunately with limited success.

Table 1 shows a rather close gender balance of 210 (47.3%) female users and 234 (52.7%) male users, as well as near-

identical turn averages for female and male users, i.e. 30 and 29, respectively.

Item counted	Totals
No. conversations	**498**
Age <10	**49**
Age 10-18	**240**
Age >18	**164**
Age unknown	**45**
Male	**234**
Female	**210**
Gender unknown	**54**
No. countries	**29**
No. turns all	**13739**
Av. no. turns all	**28**
No. turns <10	**1267**
Av. no. turns <10	**26**
No. turns 10-18	**7563**
Av. no. turns 10-18	**32**
No. turns >18	**4328**
Av. no. turns >18	**26**
No. turns age unknown	**581**
Av. no. turns age unknown	**13**
No turns male	**6689**
No. turns female	**6310**
Av. no. turns male/female	**29/30**
No turns gender unknown	**740**
Av. no. turns gender unknown	**14**

Table 1. Basic simulation data.

To enable analysis of the extent to which the specified first HCA prototype actually does reach its target user audience, Table 1 splits the users into three age groups: the under-10 year olds, the 10-18 year olds, and the over-18 year olds, representing approx. 10.8%, 52.9%, and 36.2 of the users who told HCA their age, respectively. The relatively low proportion of under-10 year olds may be explained by the fact that most under-10 year old users come from nations in which English is not a first language and hence do not speak English well enough (yet) to engage HCA in conversation. The top-five nationalities in per cent of those who told HCA their nationality, are: Denmark (28.3%), The Netherlands (15.2%), Sweden (11.3%), Norway (9.2%), and Germany (6.7%). The first nation having English as first language is the USA in 6th place (5%). In conformance with the explanation above, we find a higher proportion of speakers from countries having English as first language among the under-10 year olds, i.e. 14/40=35.0%, than of speakers from English speaking countries in proportion to all speakers of known nationality, i.e. 54/434=12.4%.

REACHING THE TARGET USERS

Let us define a turn-level metrics called *relative target group success* (RTGS) in order to quantify how well the simulated application manages to appeal to its target users as compared with its appeal to other user groups. Since the application is designed for edutainment, we consider length of conversation a component measure of success: the longer a user wants to talk to the system, the more successful is the system in meeting its edutainment objectives. We therefore propose to initially measure target group success as the percentage difference between average turn length for the target group and for each of the non-target user groups, i.e.:

$$\text{RTGS} = \frac{\text{TG-OG(n)}}{\text{OG(n)}}\,\%$$

where TG is the target group and OG(n) is some non-targetted user group.

Although we will be applying the metrics to target and other *age* groups, the metrics itself is independent of group definition. It may just as well be applied to, e.g., male and female users of an application targetted at female users.

For the three age groups, i.e. the <10, 10-18, and >18 year olds, the average turn number is 26, 32, and 26, respectively (Table 1). Thus, overall, the target user group conversations are, on average, 23.1% longer than the conversations with both non-target user groups. However, before considering this result an authoritative measure of RTGS, we need to consider the following problem.

	NNE <10	**NNE 10-18**	**NE <10**	**NE 10-18**	**NE >18**	**NGE <19**
No. users	26	203	14	23	17	29
No. turns	670	6396	514	878	468	1019
Av. no. turns	**26**	**32**	**37**	**38**	**28**	**35**

Table 2. Speaker origins. NNE is non-native English speakers, NE is native English speakers, NGE is native or good English speakers.

The 226 10-18 years old users with known nationality in the corpus are mostly non-native English speakers. Only 10.2% (=23) are native English speakers by country, i.e. come from countries which have English as a first language (Table 2). The rest, i.e. 89.8%, may be presumed to be in the process of learning English as a second language. These users are likely to be less articulate than native English speakers in conversation with HCA. We hypothesise that they might therefore tend to stop the conversation earlier than they would have done had their English been more fluent. This would make it difficult for them to match the turn average of their native English speaking counterparts of the same age. The hypothesis, thus, is that the simulated application may well have a higher-than-23% RTGS since most target users may have had a somewhat briefer conversation with HCA than they would have had, had their English skills been more mature.

To test the hypothesis, let us first compare the turn averages of the 10-18 years old native and non-native English speakers (-by-nation). Table 2 shows that the native English speaking target users have a considerably higher turn average, i.e. 38, than the non-native English speaking target users whose turn average, i.e. 32, is the same as the one for all 10-18 years old users (Table 1). This effect of mastering the English language is confirmed when we look at the turn averages of the <10 year olds. The native English speaking kids have a turn average of 37 whereas their non-native counterparts are down at 26 turns. To control for the possibility that mastery of English could be the key factor in making users speak longer with HCA, we may compare the turn average for native English speaking target users with that of native English speaking adults. Table 2 shows that the native English speaking adults had 28-turn conversations with HCA on average. This is only two turns, or 7.7%, above the average number of turns for adults in general (Table 1), showing that, although English mastery may have an effect on the length of user-HCA conversations, this effect is far smaller than the effect of belonging to the target user group. As a final test of the hypothesis of the effect of language mastery on RTGS, we may consult the linguistic grading of the English proficiency of all users on a four-point scale from bad through medium to good and native. Table 2 shows the turn average of all <19 years old native and good English speakers from Day 1 through Day 5 in the corpus. The average of 35 would seem to smoothly fit the hypothesis that, the better the English of the target users, the higher their turn average.

In conclusion, whether or not a user is in the target age group, the better the user's English skills, the longer that user is likely to speak with HCA up to 38 turns on average per conversation. Considering native English speakers only, the <10/10-18 RTGS is only 2.7% whereas the 10-18/>18 RTGS is 35.7%. These figures are +/- an estimated factor <0.1 since approx. 10% of all users did not tell HCA their age and/or nationality and since those users had far briefer conversations with him.

The marked RTGS difference just described between, on the one hand, the <10/10-18 years old and, on the other, the 10-18/>18 years old, suggests that the application clearly has stronger appeal to the <19 years old than to adults. This conclusion is supported by another finding, i.e. that the top-ten user-HCA conversations, which have a staggering average of 111 turns, all involve 6-17 years old youngsters.

CONCLUSION

This paper has proposed a simple turn-level metrics called relative target group success (RTGS) for quantifying how well an edutainment or entertainment application manages to appeal to its target users. The metrics were then applied to a relatively large (13.739 turns) WoZ corpus. It was shown that the RTGS was highly dependent on whether the defined user groups could or could not be assumed to have English as a first language. This led to the conclusion that RTGS must be measured for native speakers.

Assuming significant numbers of native English speakers in future field tests of the system, the RTGS metrics can be used directly for progress evaluation. However, even in the absence of significant numbers of native speakers, we might use the figures reported above heuristically as incremental constants. We have seen (Table 2) that: native English speaking <10 year olds talk 42.3% longer with HCA than their non-native English speaking counterparts; 10-18 years old native speakers talk 18.8% longer with HCA; and native speaking adults talk 7.7% longer with HCA than all adults (Table 1). In the absence of hard data on, e.g., <19 years old native speakers, we might compute the <10/10-18 years old RTGS for the application using non-native data as:

$$\text{RTGS TG:<10} = \frac{(\text{TG}+18.8\%)-(<10+42.3\%)}{<10+42.3\%}\,\%$$

The more future test <10/10-18/>18 non-native English turn average proportions mirror those found in the WoZ corpus, the more reliable this heuristics might be.

We obviously aim to maximise TG/non-TGs RTGSs in future work, especially the TG/adult RTGS. However, we have no idea of what might be a satisfactory RTGS in absolute terms. In fact, this question may be undecidable. A hard question which does require an answer, on the other hand, concerns *absolute* entertainment success evaluation. For instance, does an average of 38 turns (Table 2) demonstrate edutainment success in absolute terms? If not, how high must the average be? We hope that the upcoming controlled target user test with the first HCA prototype will provide part of the answer, among other things because that test will allow us to interview the target users, something which is notoriously difficult to do in field trials such as the one reported above.

When applying the RTGS metrics, care must of course be taken to exclude other possible factors. In the present case, e.g., wizard differences do not seem to influence the results.

ACKNOWLEDGMENTS

The NICE HCA work is being supported by the EU Human Language Technologies programme under contract IST-2001-35293. We gratefully acknowledge the support. Thanks are also due to the excellent work of the wizards and to Thomas Hansen who graded the users' English language proficiency.

REFERENCES

1. Bernsen, N.O., Dybkjær, H. and Dybkjær, L. *Designing Interactive Speech Systems. From First Ideas to User Testing.* London, Springer Verlag, 1998.
2. NICE: http://www.niceproject.com/
3. Walker, M., Kamm, C., and Litman, D. Towards developing general models of usability with PARADISE. *Nat. Lang. Engineering* 6, 3, 2000.