Using a Wizard of Oz as a baseline to determine which system architecture is the best for a spoken language translation system

Marianne Starlander

ETI-TIM-ISSCO, University of Geneva

40, bd du Pont d'Arve, 1211 Genève 4

Marianne.Starlander@eti.unige.ch

Abstract

This paper is part of an extended study on system architectures, the long term aim being to determine if a unidirectional, a bidirectional or a fixed-phrase architecture is more suitable in the context of the spoken language translator in the medical domain (MedSLT). Our aim here is to compare data collected during a Wizard of Oz (WOz) experiment with data collected using a beta bidirectional version of our system.

1 Introduction

The most common architectures for a spoken language translation (SLT) system are unidirectional, bidirectional or fixed-phrase systems. Unlike most commercial SLT systems for medical diagnosis, MedSLT is grammar-based. The aim is to provide reliable translations of the doctor/patient interview (Bouillon et al., 2005) in a context of controlled dialogue. In this domain precision is more important than robustness for out-of-coverage sentences since the medical user will be trained with the coverage before using the system. At first we implemented a unidirectional version because most doctor-patient interviews are doctor-initiated. However, the demand for a bidirectional system has grown and we decided to start to build such a system, but the question of which system is really best suited for such a task remained open. The aim of this study is to collect evidence to justify the choice of building a bidirectional system.

We will describe the experiments (section 2) we carried out and the resulting evaluation (section 3), before concluding in section 4.

2 Experiment

In the first phase, we constructed a WOz experiment where the participants used three different architectures (bidirectional, unidirectional and fixed-phrase) inspired from the actual MedSLT system. The users were then asked to answer a usability questionnaire where they conclude by citing their preferred architecture. In a second phase, once the actual bidirectional system was built, our aim was to conduct an experiment that would confirm the WOz user's preferences for a bidirectional system. We thus asked the same subjects to use both the beta-version of the bidirectional system and the unidirectional system and to rank the systems again according to their preference. The purpose was to check whether the constrained bidirectional system as opposed to the WOz system was still the preferred architecture. This second experiment also allowed us to study how the users adapted to a system restricted by limited coverage. In this sense, the WOz plays the role of a baseline.

2.1 WOz

Our source of inspiration was the use of a WOz experiment to collect natural data as a working basis to develop the Spoken Language Translator in the ATIS domain (Bretan et al., 2000). This type of experiment is often used for the purpose of developing a spoken dialogue system because it enables (1) the collection of representative speech data and (2) the observation of human-computer interaction in order to improve or create the interface design (Life et al., 1996). In our case, the aim is to enable users to experiment with different architectures of a system in a WOz setting. This experiment also gave us the opportunity to observe the natural interaction of doctor-patient users if they were not restricted to limited coverage. Our experimental environment was simple: the computer running the simulated MedSLT system by the doctor and patient was connected through a VNC connection to two computers in a separate room where two wizards were in reality recognizing and translating instead of MedSLT. The users were not aware that the system was actually run by humans.

2.2 Beta bidirectional MedSLT

MedSLT's bidirectional version works in a manner similar to the unidirectional version: recognition and translation is based on general unification grammars written in the Regulus format (Rayner et al., 2006). The new part is the integration of a second system for the treatment of answers. These are currently limited to elliptical sentences directly related to the question asked, so that the same ellipsis resolution can be applied to them (Bouillon et al., 2007). In order to compensate for the fact that the coverage is quite restricted due to this grammar-based approach, we provide the user with a help module that guides them towards the correct formulation. This module simply uses the result of the secondary statistical recognition to derive a list of in-coverage sentences.

For this second phase of the experiment the major challenge was to find an efficient way of training the users with the real system without interfering too much with their natural interaction with it. This training included four steps for the doctor: (1) learning the interface and the mechanical use (e.g. clicking before talking), (2) learning how to formulate questions through given controlled language rules (derived from the observations made during the WOz experiment), (3) reading through a list of in-coverage sentences during a limited amount of time, and (4) by testing the system with a member of a team to check that the microphone position and the basic usage of the system is adequate. For the patient training the main rule to observe was to answer with elliptical sentences.

2.3 Set-up

In both cases, the task was the following: the doctor or the final year medical student had to make a diagnosis for a patient who only spoke Spanish. The patients were native-Spanish speakers who were asked to pretend not to understand any French or English if they happened to do so, and to simulate sore-throat symptoms described in the task scenario. The doctor had to determine whether they suffered from a strep throat or a viral sore-throat.

In the WOz, we had eight patient-doctor pairs, each using the three different architecture versions (unidirectional, bidirectional and fixed-phrase) varying between the headache and sore-throat domains. For the actual system, three of the same doctors participated and interviewed five out of the eight original patients, and each interviewed two to three patients during a session using first the bidirectional and then the unidirectional system.

At the end of each diagnosis, lasting between ten and fifteen minutes for the real system and fifteen to thirty minutes with the WOz, the doctors filled out a diagnosis form to check on the completion of the task. In the end both doctors and patients filled in a questionnaire. This data plays a key role in the evaluation we will now describe.

3 Evaluation

We follow the classical divide in our evaluation between objective and subjective data. In the first category we decided not to include WER and SER as these measures are not really very efficient to judge the quality of a SLT system (Wang et al., 2003). Instead of WER and SER, we checked the percentage of sentences correctly translated by the system and those that were out of coverage, as this is the most important in order to guarantee an efficient doctor-patient communication. We kept the following usual measures in SLT evaluation campaigns (Stallard, 2000): task completion, and duration. We also decided to carry out a close analysis of the collected speech data regarding the type of answer formulation used by the patients. Finally in the subjective evaluation category we used a utility questionnaire.

3.1 Translation quality and task completion

In this section we will briefly comment on the quality of the translation with the bidirectional system: we divided the collected data into well translated (68.5%), badly translated (0.5%) and out-of-coverage sentences (31%). It is important to note that although the rate of out-of-coverage (OOC) sentences is quite high - it still remains clearly under the WOz OOC percentage of 74.1% - this did not affect efficiency as the average duration of a diagnosis was 12.57 minutes (compared to 20.72

min. for the WOz), and the percentage of successful task completion was around 72%. However it is important to note that this rate would be even higher if our patients really suffered from these symptoms. Patients indeed sometimes gave the doctors incoherent information, not written in their scenario, which explains most of the diagnosis errors.

3.2 Data analysis

As we were beginning to build the bidirectional version of the system, we wanted to have data about the types of answers a patient would give in response to diagnosis questions, in order to gather information on how well the users can adapt to a more limited coverage.

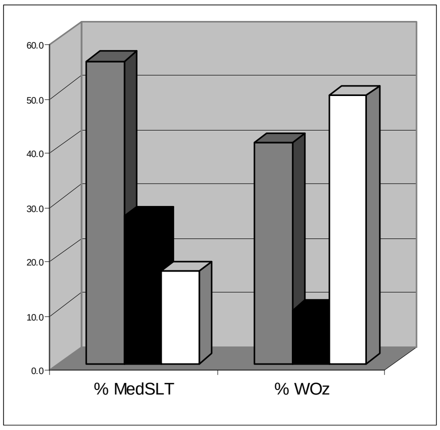

Figure 1. Ellipsis use with System X and WOz

For this reason, we specifically analyzed the proportion of ellipsis, compared to full sentences and yes/no answers. Figure 1 gives a synthesis of this study (grey = ellipsis, black = yes/no and white = full sentences).

From Figure 1 we can draw the following conclusions. First, the patient could adapt to the use of ellipsis, as shown by the fact that they used full sentences only 17.1% of the time while this percentage was far higher in the WOz. It is interesting to note that the gap in ellipsis use between WOz and MedSLT is not as wide as expected (55.7% vs. 40.7%). This would tend to prove that the use of ellipsis is quite natural when answering certain questions (e.g.: temporal questions « *Desde cuándo le duele la garganta* » - *For how long have you had your sore throat*). While questions about the location (*where is your pain?)* and the nature of symptoms (*do you have a rash?)* seem to be answered more naturally with full sentences (*sí, tengo una erupción cutánea* » - *yes, I have a rash*). Finally, patients answer much less frequently with yes/no in the WOz, since the doctor can ask more open questions like « so, what is the problem » than with the actual bidirectional system where the secondary symptoms had to be enumerated, which explains why 27.1% of the answers are of the yes/no type.

3.3 Questionnaire

Based on (Lewis, 1991) we constructed a usability questionnaire, using a 1-5 Likert scale to grade the answers given to the following questions :

Q1	Easy to use the system
Q2	Clear instructions on task
Q3	Good response time
Q4	Could ask enough questions to be sure of diagnosis
Q5	System more efficient than non-verbal communication
Q6	User-friendly interface
Q7	Utility of CL rules
Q8	Help window very useful to learn coverage
Q9	Have often taken sentences directly from help window

Table 2. Abstract of questions

Figure 2 synthesizes the answers to the questionnaire. The real system obtains higher scores than the WOz for all questions apart from Q5, where both obtain almost equally high results. This tells us that both systems are more efficient than non-verbal communication. The less differentiated scores for Q4 are due to simulation of symptoms which sometimes made patients answer in a less clear-cut manner which sometimes puzzled the doctors. This probably explains why the score is no higher than 4 for MedSLT and 3.7 for the WOz. Interestingly MedSLT gets higher scores. This would definitely tend to prove that the constraints due to limited coverage were not impeding the dialogue interview. The most important gap between MedSLT (4.3) and the WOz (2.1) is quite logically found in the question about the speed of the system (Q3). The results for Q1 show that the participants declare that they could easily learn how to use the system thanks to the given instructions. Interestingly, the gap between the WOz and the real system is not wider, as we would have expected since the users have to adapt to the limited coverage.

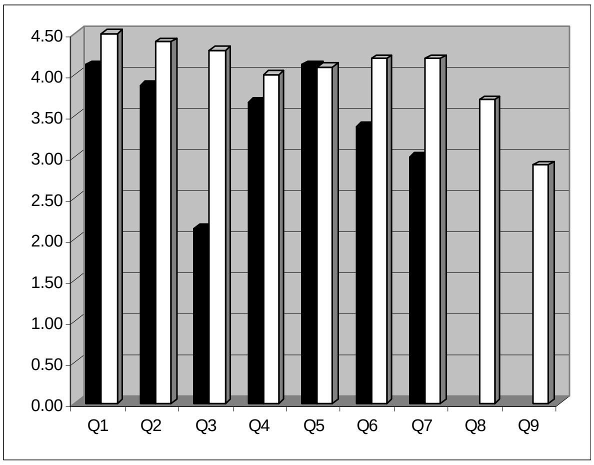

Figure 2. Answers to the questionnaire (white=our system, black=WOz)

This leads us to the results for Q7-Q9 that all concern the learning of the system's coverage. The real system scores high on Q7 about the utility of the controlled language (CL) rules which were given in order to guide the user's formulation. However, the participants gave mitigated answers about the utility of such rules, which can be explained by the unrestricted nature of the WOz. The CL rules were considered to be more useful to learn the coverage than the sentences displayed in the help window (Q 7), whereas these were deemed very useful in previous studies (Starlander et al., 2005).

Finally, the questionnaires tell us that when comparing the different available architectures, the users always prefer a bidirectional architecture, even with the beta version of MedSlt where the coverage is more restricted.

4 Conclusion

After this study using a WOz as a baseline system we can conclude that the bidirectional MedSLT system is performing well; and that the users still prefer this architecture. The users, especially the patients, can adapt to its limited coverage, by using ellipsis and thus achieving a very acceptable task completion. The overall translation quality is acceptable.

This work is only part of a more extended study comparing different architecture with regard to usability and user satisfaction. The next step, before an extended evaluation, involves a further development phase, after which we would like to compare the actual restricted version of the bidirectional system with a wider version allowing full sentences in some extent.

References

Bouillon, P., M. Rayner, et al. (2005). A generic Multlingual Open Source Platform for Limited-Domain Medical Speech Translation, *10th Conference of the European Association of Machine Translation*, Budapest, Hungary: 50-58.

Bouillon, P., M. Rayner, et al. (2007). Les ellipses dans un système de Traduction Automatique de la Parole, to appear in *TALN 2007*, Toulouse.

Bretan, I., R. Eklund, et al. (2000). Corpora and Data Collection. *The Spoken Language Translator.* M. Rayner, D. Carter, P. Bouillon, V. Digalakis et M. Wirén. Cambridge, Cambridge University Press: 131-144.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*: 189-194.

Life, A., I. Salter, et al. (1996). Data collection for the MASK kiosk: WOz vs prototype system. *ICSLP'96*, Philadelphia, USA: 1672-1675.

Lewis, J. R. (1991). "Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ " *SIGCHI Bulletin* 23(1): 78-81.

Rayner, M., B. A. Hockey, et al. (2006). *Putting Linguistics into Speech Recognition.* Stanford, California, Stanford University Center for the Study of Language and Information.

Stallard, D. (2000). Evaluation Results for the Talk'n'travel System. *Applied Natural Language Processing Conference, Seattle*, Washington.

Starlander, M., P. Bouillon, et al. (2005). Practising Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT). MT Summit X, Phuket, Thailand: 188-194.

Wang, Y.-Y., A. Acero, et al. (2003). Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. *Workshop on Automatic Speech Recognition and Understanding*, St Thomas, US Virgin Islands.