

Comparing Intrinsic and Extrinsic Evaluation of MT Output in a Dialogue System

Anne H. Schneider, Ielka van der Sluis, Saturnino Luz

Department of Computer Science
Trinity College Dublin, Ireland

`schneia@scss.tcd.ie`

Abstract

We present an exploratory study to assess machine translation output for application in a dialogue system using an intrinsic and an extrinsic evaluation method. For the intrinsic evaluation we developed an annotation scheme to determine the quality of the translated utterances in isolation. For the extrinsic evaluation we employed the Wizard of Oz technique to assess the quality of the translations in the context of a dialogue application. Results differ and we discuss the possible reasons for this outcome.

1. Introduction

In the last two decades machine translation (MT) technology has reached a level of quality, which warrants its increasing use in real-world applications. A significant trend in this direction has been the use of MT in combination with speech technologies. GOOGLE, for instance, has recently announced that it is working on speech to speech translation software for mobile phones. However, machine translated content is still far from perfect. For MT components to be effectively deployed, it is important for developers to be able to reliably assess the quality of the translation output. This task is difficult for a number of reasons [1]. MT has traditionally been evaluated *intrinsically*, that is, independently of the system in which it will be used, or the task to which the system will be put. This is done through comparable evaluation metrics, which aim to correlate as closely as possible to human judgement performed on a task independent (and often sentence-by-sentence) basis. Leaving aside the question of whether a combination of such idealised judgements truly represents the perceived quality of a translation, finding an objective metric poses considerable challenges. The naive approach of simply counting mistranslated words rarely produces meaningful results. Often there is no single “correct” translation. A text can have several acceptable translations, and a quality rating in this case is more a matter of taste than of what is right or wrong. Furthermore, the boundaries of errors are hard to determine. Errors frequently involve whole phrases and even discontinuous expressions. As one error can lead to another, the cause of an error is not always apparent, which hinders the discovery of what went wrong in the translation.

Although most MT evaluation is still intrinsic, there have been renewed calls for more attention to be paid to *extrinsic*

evaluation in MT and NLP in general [2, 3]. Extrinsic evaluation aims to assess the (often indirect) effect of a module, such as an MT component, on task- and context-dependent variables such as user performance, through its performance as part of a functioning system. The main difficulty with this kind of evaluation is the effort usually required to build a system to be tested by users. In addition, the issue of how to combine these different types of evaluation needs to be investigated.

This paper reports on an exploratory study that employs two types of evaluation, an intrinsic and an extrinsic method to examine MT output for use in a dialogue system, to aid applications using MT combined with speech technologies. Text used in a dialogue differs considerably from written text [4]. Dialogues are, for example, more interactive and contain direct references (i.e. to the addressee). Language production in dialogue is prompt whereas in written text typically more consideration is put into the formulation. Even in situations which deviate from natural, face-to-face dialogue settings, such as spoken-language dialogues between remotely located participants or interactive, time-constrained situations where different modalities are employed, language use varies with respect to standard written text. It is therefore not obvious that a machine translation system that performs well on written text will also yield good usability results in interactive systems. Similarly, poor machine translation performance in text does not necessarily imply poor usability in interactive contexts.

We chose a dialogue setting for the experiment because in human computer interaction spoken dialogue offers an immediate and human-like means of communication that suits many applications in for instance hands-busy, eyes-busy situations. In addition, the state of the art in spoken dialogue systems has grown to a point where such applications are feasible. A simple prototyping method was used to collect the set of English output utterances for our system. Subsequently, two web-based, state-of-the-art MT systems were employed to obtain German translations. For the intrinsic evaluation, in which we assessed the quality of the translated sentences in isolation, three human judges were asked to indicate which translation they preferred. An annotation scheme was developed to investigate the reasons behind the preferences expressed by the judges. For the extrinsic evaluation the translations were incorporated into a Wizard of

Oz (WOZ) system, and an experiment was run with eight German subjects who interacted with the system. Results of these interactions were gathered using questionnaires, interviews and system logs, as well as through an analysis of the small corpus of dialogues collected with the experiment.

2. Related work

2.1. MT Evaluation

Currently, MT evaluation mostly employs automatic metrics such as BLEU [5], NIST [6], and METEOR [7] to enable researchers to validate and optimise translation methods quickly [8]. To assess the translation quality BLEU, and its variant NIST, count the number of n-grams, of varying length, of the system output that can also be found in a set of references. METEOR, on the other hand, measures the number of word matches between the output of the system and the reference. In a second step all unmatched words get stemmed and matched with the reference again. Reordering of words is penalised. Another automatic evaluation method is TER which has been introduced by the Global Autonomous Language Exploitation (GALE) research program. TER calculates the minimum number of edits that are needed to change a hypothesis so that it exactly matches the closest reference, from a set of references. This number is normalised by the average length of the references. Edits can be insertions, deletions, substitutions and shifts. For the human-in-the-loop alternative HTER [9] human annotators produce the closest reference. The TER score is then automatically computed on basis of this human corrected reference. These quality scores either assume large corpora of text or rely on a large number of references in order to correlate with human judgements. Therefore it is not meaningful to calculate one of these scores for the small set of dialogue utterances that were used in our study.

Some intrinsic evaluations related to the work presented in this paper have been reported by Turian et al. [10] who asked human judges for ratings of MT output before the development of automatic evaluation methods. Also, Kit et al. [11] report on a comparative evaluation of the BLEU and NIST scores of six representative online MT systems including SYSTRAN and GOOGLE for translating legal texts from various languages into English. These intrinsic methods are ‘decontextualized’ in that they do not take into account the task being supported by MT, and therefore might not produce meaningful results in real-life situations [12]. In such applications time constraints, distractions, accommodation to certain types of mistakes by the user, awareness of translation errors in spoken language as opposed to text, and other issues that arise in interactive situations might influence the perception of MT output.

Only a few, isolated extrinsic evaluation efforts, which assess the effect of task- and context-dependent variables on the user performance of MT systems can be found in the literature. Most of these efforts date back to the late 90s, when

two new MT research trends emerged [13]. One was task-based experiments conducted by developers of MT systems. Levin et al. [14], for example designed a task based evaluation method for the JANUS speech-to-speech MT system and compared the results of this evaluation with their accuracy based evaluations and Phillip Resnik [15] proposed a method to evaluate multilingual gisting based on its role of decision support. The second evaluation stream was task-based experiments assuming an ordering of task difficulty for text-handling tasks such as proposed by Taylor and White [16] and White et al. [17]. In 2006 the idea of task based evaluation was picked up again by a research group around Laoudi and Voss who conducted experiments on the text-handling task of extracting information from MT output [13, 18].

2.2. Wizard of Oz studies

The Wizard of Oz technique is an early stage prototyping method by which a system can be tested without first having to build it [19]. As in the novel *The wonderful wizard of Oz*, from which the technique got its name, a user is presented with what appears to be a working system, while a human operator (the “wizard”), who is not visible to the user, takes the role of the system. The method is a powerful technique when the input modality has a high computation/cognition ratio in the sense that it can be only partially decoded by computers but is easily understood by humans [20]. Therefore, the WOZ technique has a long tradition in the design of speech systems [21]. Studies where the WOZ technique was used in the context of speech applications include: the collection of corpora to tune speech recognizers [22], the evaluation of the perception of different input modalities [23], and the analysis of speech based design ideas, such as the smart home [24] and VICO, the driving assistant [25]. Whittaker et al. [26] present a study where a wizard is used to discover dialogue strategies in the restaurant domain. In the context of machine translation, studies that use WOZ are rare. In the Verbmobil project, which aimed at developing a system capable of interpreting dialogues, the method was employed to explore strategies and phenomena in interpretation [27]. Apart from that we are not aware of any other studies where the WOZ technique is used to evaluate or discover machine translation output.

3. Study

3.1. Material

Our system means to facilitate a scenario in which German speakers have to find a good offer on Internet connections in Ireland. The knowledge acquisition for this system was done through inspection of the web sites of Irish Internet providers. To collect the necessary system utterances we simulated human-computer dialogues using a regular chat tool, which proved to be a simple yet adequate prototyping approach to test the set of system utterances for completeness. Five participants were asked to use the chat tool to find

out about good offers on Internet connections in Ireland. Our experimenter ‘helped’ them using a fixed set of predefined system utterances. After each dialogue the set of utterances was adjusted depending on what appeared to be missing. The dialogue length was usually around 20 turns, but the number of words within the conversations ranged from less than 200 to more than 500. One of these more complex dialogues was caused by a person ‘playing around with the system’ to figure out its limitations. Another dialogue ended in a breakdown-like situation in which the set of utterances was not sufficient and the experimenter had to create new utterances. Our final set of system outputs consisted of 32 utterances (10 of which had open slots) and a list of slot fillers. The 32 English utterances consisted of 1 welcome message, 3 utterances that signal a problem in the interaction and suggest how to proceed, 6 explicit feedbacks to user input, 4 information requests, 15 utterances that provide information on a particular option, 1 stalling, 1 break off, and 1 goodbye message.

This set of utterances was, firstly, translated to German by a native German speaker. We call this human translation our reference translation. Secondly, the English source utterances were translated using the online machine translation service of SYSTRAN (<http://www.systranet.com/>) and GOOGLE (<http://translate.google.com/>).¹ In this paper we are interested in the quality of the two machine translations.

3.2. Method

Inspired by the paper by Belz [3], who addresses the issue that off-line, intrinsic evaluations do not necessarily result in useful applications, we performed a two-fold evaluation of the quality of the machine translations of the system utterances: (1) an intrinsic evaluation in which the quality of the utterances was assessed in isolation and (2) an extrinsic evaluation in which the quality of the utterances was assessed in their context of use (ie. our dialogue application). Our hypothesis was that the translation with the best ratings in the intrinsic evaluation would produce the smoothest interaction in the extrinsic evaluation.

4. Intrinsic evaluation

Materials: Four sets of 28 system utterances, namely, the English original, the German reference translation, the SYSTRAN translation and the GOOGLE translation. Note that we excluded the list of slot fillers and the utterances that rendered identical translations by GOOGLE and SYSTRAN.

Subjects: Three human judges, native German speakers with language experience in English speaking countries. None of the judges was involved in the collection of the utterances or saw them before the evaluation.

Procedure: As discussed in section 2, existing MT eval-

¹All automatic translations have been produced on November 7th 2009. GOOGLE’S translation service is subject to constant change, therefore it is possible that reported translation errors do not appear at a later point in time.

uation methods are not meant for small sets of dialogue utterances that considerably vary in their syntax and pragmatics and that are meant to be used in a particular context. Instead, we decided to do the intrinsic evaluation in a way that preserves some of the characteristics of the in section 2 described automatic metrics but relies on human judges. For each of the 28 system utterances our judges were shown the English original and the two machine translations. They were asked to identify which of the two translations they preferred and to point out the mistakes in the less preferred translation, which induced their judgment.

Results: An overview of the preference rating (in numbers) and the agreement (Kappa scores) is presented in Table 1 and shows a high agreement between our judges. In almost two thirds of the cases, the GOOGLE translation was preferred over the SYSTRAN translation.

	GOOGLE	SYSTRAN		Kappa
j1	20	8	j1 + j2	0.6725
j2	18	10	j1 + j3	0.8444
j3	18	10	j2 + j3	0.8363

Table 1: GOOGLE and SYSTRAN preferences of judge j1, j2 and j3. Cohen’s Kappa Score for three pairs of judges.

Further Analysis: To understand the reasons for the preferences of our judges and to gain a better insight in the translation errors we performed a further analysis of the automatic translations against the human reference translation similar to TER. First, we calculated the edit distances between each of the two automatic translations and the human reference translation by summing the number of added words (ie. words that do not appear in the reference translation) and the number of deleted words (i.e. words that do not appear in the machine translations but that do appear in the reference translation). However, results of this exercise presented at the top of Table 2 show no difference between the two machine translations.

We proceeded by developing an annotation scheme that covers translation errors like:

- *wrong word order*: the order of words in the translated utterance does not correspond to the order of words in the utterance in the reference;
- *wrong word*: word which does not appear in the reference and which is not a synonym;
- *synonym*: noun or verb which is not the same as the word in the reference but which has a similar meaning;
- *untranslated word*: word from the source utterance, which has not been translated.

The same three judges that we asked before were now asked to compare both the SYSTRAN and GOOGLE translations with our reference using our annotation scheme.

4.1. Results

To find out why our judges preferred the GOOGLE translation over the SYSTRAN translation, we started with making a sim-

ple correction of the edit distance. In calculating the edit distance, each synonym causes a deletion and an addition while the meaning of the utterance stays the same. Hence, we adjusted the edit distance calculations by taking into account the number of synonyms (ie. edit distance minus 2 times the number of synonyms). Table 2 shows that GOOGLE and SYSTRAN perform similarly when taking into account the number of synonyms.

	GOOGLE			SYSTRAN		
add	110			111		
del	116			115		
edit	226			226		
	A1	A2	A3	A1	A2	A3
syn	25	23	23	27	20	27
adedit	176	180	180	172	186	172
wwo	3	3	3	4	4	4
ww	31	49	21	64	80	21
untr	10	11	11	0	0	0

Table 2: GOOGLE and SYSTRAN comparison including *edit* distance, *added* words, *deleted* words, adjusted edit distance (*adedit*), *synonyms*, the number of wrong word orders (*wwo*), wrong words (*ww*) and untranslated words (*untr*) identified by annotators A1, A2 and A3.

We proceeded with a further examination of the annotations. Results in Table 2 indicate that (1) the use of wrong word orders was similarly distributed in the output of the two translation systems; (2) GOOGLE leaves some words untranslated, while SYSTRAN appears to translate everything; and (3) compared to the GOOGLE translation, annotators A1 and A2 found almost double the amount of wrong words in the SYSTRAN translation.

When looking more closely at the untranslated words in the GOOGLE output, we notice that the complete goodbye phrase ‘*Have a good day.*’ stayed untranslated. Another example is the word ‘*sorry*’ at the beginning of an phrase that signals a problem in the interaction. The use of the word ‘*sorry*’ has become very common in the German language, so it is unclear if the word is really untranslated or if the actual translation is the word ‘*sorry*’. But the key to the judges’ preference for the GOOGLE translation must be in the number of wrong words that appeared in the translations. Our data shows that in this category GOOGLE performs much better than SYSTRAN. Note, however, that in this category annotators do not agree very well, although they had clear instructions of what determines a wrong word.

We concluded that we would need a more detailed classification of wrong words and identified three further subgroups. Wrong words can be the result of the fact that the translators choose the wrong meaning for an ambiguous word in the source. The SYSTRAN system for example used the German word for unhappy (‘*traurig*’) to translate ‘*sorry*’, which was intended as an excuse and therefore should rather be translated as ‘*Entschuldigung*’. Spelling mistakes can also cause wrong words. We found an example

of this in the SYSTRAN translation. The translation for ‘*internet connection*’ ended up as ‘*Internetanschlusse*’ rather than ‘*Internetanschluss*’. The third group of wrong words comprises compounds that the translation system builds up from translations of two or more source words but that can not actually be found in the German language. This last error type was unique to the SYSTRAN system. An impressive example for the creative ability of SYSTRAN is the compound of ‘*Überlandleitung*’ for landline and ‘*Verbindung*’ for connection into the word ‘*Überlandleitungverbindung*’ as a translation for ‘*landline connection*’ (‘*Festnetzanschluss*’). Another example is the compound ‘*Kilobyteantriebskraftgeschwindigkeit*’ for ‘*kilobyte upload speed*’. We would like to conclude at this point that, although our intrinsic evaluation already goes beyond simple calculations of edit distance and word error rate, a further analysis of different types of erroneous words promises to be informative in determining MT quality.

5. Extrinsic evaluation

Materials: We conducted a Wizard of Oz experiment with a prototyping tool [28], which allows for speech input and text output and produces time stamped logs for every system utterance. Text output was chosen to make sure that our subjects were not influenced by the particular voice of the system (either synthetic or human recordings). Speech input allowed us to measure physiological changes in for instance galvanic skin responses (in this paper physiological results are however not reported). The user interface includes a ‘source button’, which, when clicked, leads to the English source of the current system utterance.

Two systems, one with the SYSTRAN and one with the GOOGLE translation were implemented. We used questionnaires to capture demographic data of the participants and to assess the quality of the translations. In the latter questionnaires a 5-point Likert scale was used, ranging from 1 (strongly agree) to 5 (strongly disagree). Questionnaires, instructions etc. were presented to participants on paper. Interactions with the system and interviews with the participants were audio recorded.

Two scenarios were used in the study:

- Imagine you stay in Dublin for the period of 6 months. You are looking for an Internet connection, which you can use at home as well as on campus of your University. The offer should be as cheap as possible.
- Imagine you are looking for an Internet connection at your home, which is as fast as possible. You are also looking for the highest possible download allowance. The price of the connection therefore does not matter to you.

Subjects: Eight German ERASMUS students participated in our study. They were reasonably fluent in English, had moderate to low computer skills and only little experience with dialogue systems. Their average age was 22, ranging from 20 to 24.

Procedure: After filling out a questionnaire to obtain de-

mographic data, participants were asked to read the introduction to the study, which explained that the study was meant to try out two versions of a dialogue system that provides information on Internet offers. Subjects were also told that the dialogue systems had recently been automatically translated from English to German, such that they accept German speech as input and produce German text as output, the latter due to problems with the synthesizer. In addition, the ‘source button’ was introduced as a means to trace the source of the translation in the case of problems with the system output. Next, participants were asked to read one of the two scenarios, to find a solution with one of the two systems and to fill out a questionnaire about the interaction. This sequence was repeated for the remaining scenario and system. Scenarios and systems were equally shuffled between participants. After finishing these tasks, the experimenter explained that the participants had been interacting with a human wizard instead of a system and interviewed them about the system outputs they had received using the system logs.

5.1. Results

We analysed the performance of the two dialogue systems in terms of efficiency, quality and task success [29].

Efficiency: Efficiency results, are presented in Table 3. The elapsed time, the time between the first and the last system utterances in the dialogue, favours the GOOGLE over the SYSTRAN system, not only when summed over the participants (GOOGLE 250 seconds per interaction on average (stdev 42.42), SYSTRAN 285 seconds per interaction on average (stdev 77.72)), but also dependent on whether the system was used before or after participants interacted with the SYSTRAN system. The total number of system and user turns is higher in interactions with the GOOGLE system. With respect to the user turns, this results from the fact that in the cases where the participant started with the GOOGLE system, the number of user turns is almost doubled.

	GOOGLE		
	user turns	system turns	elapsed time
1st	75	79	17:28
2nd	38	65	15:45
sum	113	144	33:23
	SYSTRAN		
	user turns	system turns	elapsed time
1st	42	69	20:05
2nd	43	65	17:32
sum	85	134	37:37

Table 3: Number of user turns, system turns and elapsed time in minutes.

Quality: Dialogue quality was measured with the use of the source button, which was clicked 25 times in all interactions. In the interviews we discovered that in five of these cases participants did not understand the system’s output

(two times GOOGLE; three times SYSTRAN). Especially, the SYSTRAN compound ‘Überlandleitungverbindung’ caused confusion. In seven cases participants did not trust the translation (three times GOOGLE; four times SYSTRAN). Participants stated that they used the button because they hoped for more information in the English source text and they believed that sometimes information had gone missing in the translation. In the other 13 cases the button was used out of curiosity because the participant wanted to see what happens, or to understand where an awkward translation originated from (five times for GOOGLE and eight times for SYSTRAN).

To assess the dialogue quality, we also examined how often the break off utterance and the three utterances that signal an interaction problem were used in the dialogues. The break off utterance ‘Sorry, but I have no information on that topic.’ was used twice (1 time GOOGLE; 1 time SYSTRAN). The negative feedback ‘Sorry, I did not understand you. Could you say that again?’ was used in seven cases (5 times GOOGLE; 2 times SYSTRAN). The system could also offer the participant to go back a step, which happened once with the GOOGLE system. The system utterance ‘Do you want to start over again?’ did not occur in the dialogues.

Task Success: Task success was measured through eight statements in the questionnaire. Results presented in Table 4 show that participants observed differences between the systems’ speech recognisers (Q8) and favoured the GOOGLE over the SYSTRAN system. Note that we used the same wizard for all interactions. Closer inspection of the data suggests that these differences have been influenced by the order in which the systems were used, usually the second system was judged better. System responses (Q9, Q10 and Q11) were generally perceived as positive and did not differ much between systems. The interaction pace (Q12) showed some differences between subjects but not between systems. With respect to the ease of the task, Q13 showed some differences between subjects independent from the order in which the systems were used. Q14 resulted in a clear, but possibly order dependent preference for the SYSTRAN system. Expected system behaviour (Q15), was judged positively and rendered no differences between systems. Finally participants were not too keen on using the systems in the future (Q16).

Perception of MT Output: In addition to the interaction analysis we included seven statements into the questionnaire to obtain a more detailed insight in how participants perceived the MT output. Notably, the statement ‘I had serious problems understanding the German texts.’ was rated negatively for both systems (GOOGLE 4.13(0.99) and SYSTRAN 4.13(.84)). Responses to 27 out of the 56 ratings (= 8 participants x 7 items) did not show any differences between the two systems. Table 4 presents the results for these items (Q1-Q7). GOOGLE performed better in the case of questions Q1, Q3, Q4 and Q5, whereas SYSTRAN performed better for Q6 and Q7.

Insights from Interviews: Interviews with participants P1 to P8 that were carried out after they had worked through

Q		GOOGLE	SYSTRAN
1	I always knew what the system was asking me for.	2.38(.74)	2.75(1.04)
2	I had serious problems understanding the German texts.	4.13(0.99)	4.13(.84)
3	I would rate the German utterances as excellent.	3.75(1.17)	4.00(.93)
4	There were awkward words and phrases in the German dialogue.	2.63(1.06)	2.25(1.03)
5	The German utterances were fluent.	3.00(.76)	3.50(.93)
6	I would rather use the English original.	2.00(.76)	2.13(1.13)
7	I would rate the German utterances as incomprehensible.	4.00(.93)	4.13(1.13)
8	The system did always understand what I said.	3.00(1.51)	2.25(1.49)
9	The system did not give me enough information.	3.63(.74)	3.38(1.30)
10	The system gave me a lot of unnecessary information.	4.38(.74)	4.125(.84)
11	The system’s responses were appropriate.	2.13(.64)	2.38(.92)
12	The system gave me too much information in one go.	3.75(1.49)	3.63(1.41)
13	The system made it easy to find the offer that I was looking for.	2.38(.74)	2.38(.74)
14	I could quickly find what I was looking for.	2.75(.71)	2.38(.52)
15	The system responses agreed with my expectation.	2.75(.46)	2.50(.54)
16	I would consider using a similar system to find a good offer on broadband Internet.	3.75(.46)	3.75(.71)

Table 4: Means and (standard deviations) for questionnaire items.

the two scenarios, reveal some background on the participants’ judgments of the MT output. For instance, wrong word orders were recognized by all participants during the interactions, but were not perceived as big obstacles for comprehension, they only ‘spoil the overall impression’ (P3). The word ‘*sorry*’ was not recognised as non-German by P1, P5, P6, but P3, P4 and P7 said they perceived it as ‘impolite in the context of an information system’. P4 and P6 only noticed that the GOOGLE utterance (‘*Have a good day*’) was not translated to German when they were asked about it in the interview. A possible explanation for this lies in what P7 said: ‘as an ERASMUS student switching between the two languages is in the daily routine’. P2 remarked that the translation of ‘*options*’ into ‘*Wahlen*’ by SYSTRAN did not really influence the comprehensibility but ‘disturbs the fluency’. In principle, ‘*Wahl*’ is an acceptable translation for ‘*option*’ in the sense of ‘*choice*’. However, the plural (‘*Wahlen*’) can only be used for the German word ‘*Wahl*’ in the sense of ‘*election*’. The spelling mistake in the translation for *internet connection* by SYSTRAN was not recognized in the interaction by any of the subjects. A considerable number of participants also only realised the absurdity of the SYSTRAN compound ‘*Kilobyteantriebskraftgeschwindigkeit*’ during the interview. A possible reason for this may be that the word appeared in a lengthy description, which was only scanned by participants to find relevant values (P7). Another reason could be the participants’ low computer literacy. For example, P1 said ‘I read it and thought that this is just another of these computer terms’ and P7 ‘saw Kilobyte and did not really read the rest of the word’. Similar responses were observed with the translations of ‘*download allowance*’. Both systems failed with the translation (‘*download Zertifikat*’ (GOOGLE) and ‘*Downloadzulagen*’ (SYSTRAN)). P1 said he thought that it was not ‘of much importance’ to understand the word properly, because it was ‘not essential to carry on

with the conversation’.

6. Discussion and Conclusion

We have reported on an assessment of the quality of MT output. We expected that the best translation would receive the best ratings in the intrinsic evaluation, and results in the smoothest interaction in the extrinsic evaluation. However, results present a different picture. Our intrinsic evaluation showed a clear preference for the GOOGLE translation. A similar result is reported in [11] where the BLEU and NIST scores of GOOGLE are slightly better than those of SYSTRAN for most language pairs.² Annotation of the MT output indicated that the preference of our judges was most likely caused by the fact that SYSTRAN included more erroneous words in its output due to ambiguous source words, spelling mistakes and the generation of non-existent compounds. However, when the MT outputs were used in a dialogue setting, the better performance of the GOOGLE system in terms of intrinsic evaluation was not reflected in the interactive context.

In terms of efficiency, it took participants slightly longer to finish their tasks with the SYSTRAN system, but in carrying out their first task, participants needed more dialogue turns when using the GOOGLE system. Our interaction quality analysis suggest that the GOOGLE system performed less well than the SYSTRAN system, since the former resulted in dialogues containing more utterances, which indicates interaction problems. From the use of the source button we could not conclude a difference between the systems. Task success, measured through participants’ ratings of a number of questionnaire items, did not show huge differences between the two systems. The assessment of the system utterances indicated that participants had no difficulties in understand-

²This study considered the domain of legal texts and compared translations from various languages into English.

ing the system and only in a few cases preferred one system to the other. In summary, our extrinsic evaluation did not render a preference for the GOOGLE system as we expected from the outcomes of the intrinsic evaluation. In the interviews we discovered that most translation errors, especially those of the SYSTRAN system, were only recognized by a small number of participants during the interaction or not at all. This may have resulted from the computer literacy of our participants but the experiment task may also have had an effect.

Despite the fact that the extrinsic evaluation is a small-scale study in a limited domain involving a single language pair, results show a difference to the evaluation of the MT output in isolation. Depending on the context of use, the de-contextualised intrinsic methods may not provide us with the most accurate results. As future work we plan further WOZ experiments in which extra errors will be added to the MT output, in order to investigate the issue of user acceptance versus error rate in greater depth.

In summary, this paper contributes to current frameworks for performance analysis by showing that intrinsic evaluations of system output may not be informative for the quality of a dialogue system. Evaluation efforts need to keep the user and application context in mind if they aim to produce meaningful results for real-life situations. In setting up the WOZ study we also discovered that existing standardised questionnaires do not capture the system performance in terms of the interaction factors necessary to determine the quality of dialogue systems. Finally we found very valuable information through subject interviews as an additional measurement method.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

8. References

- [1] M. Flanagan, "Error Classification For MT Evaluation," in *Proceedings of Association for Machine Translation in the Americas (AMTA'94)*, 1994.
- [2] J. Goldstein, A. Lavie, C. Lin, and C. Voss, Eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- [3] A. Belz, "Thats nice...what can you do with it?" *Computational Linguistics*, vol. 35, no. 1, pp. 111–118, 2009.
- [4] D. Biber, *Variation across speech and writing*. Cambridge University Press, 1988.
- [5] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLUE: a method for automatic evaluation of machine translation," in *Proceedings of ACL'02*, 2002.
- [6] S. Strassel, M. Przybocki, K. Peterson, Z. Song, and K. Maeda, "Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction," in *Proceedings of LREC'08*, 2008.
- [7] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- [8] Y. He and A. Way, "Learning Labelled Dependencies in Machine Translation Evaluation," in *Proceedings of EAMT'09*, 2009.
- [9] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of Association for Machine Translation in the Americas (AMTA'06)*, Cambridge, Massachusetts, 2006, pp. 223–231.
- [10] J. Turian, L. Shen, and I. Melamed, "Evaluation of Machine Translation and its Evaluation," in *MT Summit IX*, New Orleans, 2003, pp. 368–393.
- [11] C. Kit and T. M. Wong, "Comparative Evaluation of Online Machine Translation Systems with Legal Texts," *Law Library Journal*, vol. 100, no. 2, pp. 299–321, 2008.
- [12] N. Karamanis, S. Luz, and G. Doherty, "Translation practice in the workplace and Machine Translation," in *Proceedings of EAMT'10*, Saint-Raphaël, France, 2010.
- [13] J. Laoudi, C. R. Tate, and C. R. Voss, "Task-based MT Evaluation: From Who/When/Where Extraction to Event Understanding," in *Proceedings of LREC'06*, vol. 6, Genoa, Italy, 2006, pp. 2048–2053.
- [14] L. Levin, B. Bartlog, A. Llitjos, D. Gates, A. Lavie, D. Wallace, T. Watanabe, and M. Woszczyna, "Lessons Learned from a Task-Based Evaluation of Speech-to-Speech Machine Translation," in *In Proceedings of LREC'00*, Athens, Greece, 2000.
- [15] P. Resnik, "Evaluating Multilingual Gisting of Web Pages," in *Proceedings of the AAAI Symposium on Natural Language Processing for the World Wide Web*, Stanford, 1997.
- [16] K. Taylor and J. White, "Predicting What MT is Good for: User Judgments and Task Performance," in *Proceedings of Association for Machine Translation in the Americas (AMTA'98)*. Springer Berlin, 1998, pp. 364–373.

- [17] J. S. White, J. B. Doyon, and S. W. Talbott, "Task Tolerance of MT Output in Integrated Text Processes," in *Proceedings of Workshop: Embedded Machine Translation Systems ANLP/NAACL*, Seattle, WA, 2000, pp. 9–16.
- [18] C. R. Voss and C. R. Tate, "Task-based Evaluation of Machine Translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output," in *Proceedings of EAMT'06*, Oslo, Norway, 2006.
- [19] X. Faulkner, *Usability Engineering*. Palgrave Macmillan, 2000.
- [20] H. Dybkjaer, N. Bernsen, and L. Dybkjaer, "Wizard-of-Oz and the trade off between naturalness and recognizer constraints," in *Proceedings of Eurospeech'93*, 1993.
- [21] J. Gould and C. Lewis, "Designing for Usability- Key Principles and What Designers Think," in *Proceedings of CHI'83*, 1983.
- [22] G. Fabbriozio, G. Tur, and D. Hakkani-Tür, "Automated Wizard-of-Oz for Spoken Dialogue Systems," in *Proceedings of Interspeech'05*, 2005.
- [23] M. Rajman, M. Ailomaa, A. Lisowska, M. Melichar, and S. Armstrong, "Extending the Wizard of Oz Methodology for Language-enabled Multimodal Systems," in *Proceedings of LREC'06*, 2006.
- [24] F. Gödde, S. Möller, K. Engelbrecht, C. Kühnel, R. Schleicher, A. Naumann, and M. Wolters, "Study of a Speech-based Smart Home System with Older Users," in *Proceedings of Workshop on Intelligent UIs for Ambient Assisted Living*, 2008.
- [25] P. Geutner, F. Steffens, and D. Manstetten, "Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz Experiments," in *Proceedings of LREC'02*, 2002.
- [26] S. Whittaker, M. Walker, and J. Moore, "Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain," in *Proceedings of LREC'02*, 2002.
- [27] D. Krause, "Using an Interpretation System - Some Observations in Hidden Operator Simulations of 'VERB-MOBIL'," in *Proceedings of Dialogue Processing in Spoken Language Systems*, 1997.
- [28] S. Schlögl, G. Doherty, and S. Luz, "WebWOZ : A Wizard of Oz Prototyping Framework," 2010.
- [29] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with PARADISE," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 363–377, 2000.