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

Anne H. Schneider Ielka van der Sluis Saturnino Luz
Trinity College Dublin

further details: Schneider et al., at IWSLT 2010

Intrinsic Evaluation
- de-contextualised evaluation
- independent of system, task, and user
- developer oriented

Extrinsic Evaluation
- evaluation in context
- with the system, task and user in mind
- user oriented

-> evaluation of Machine Translation (MT) is traditionally intrinsic (e.g. with automatic metrics such as BLEU, METEOR, ...)
But: MT output is usually used in a particular context

Exploratory Study to Compare Intrinsic and Extrinsic Evaluation of MT Output

- design of system-output (English) for a dialogue system to recommend broadband Internet offers (28 utterances)
- three German translations: 1) human gold standard 2) Google translator `(translate.google.de)` 3) Systran translator `(systranet.com)`

Intrinsic Evaluation

with 3 human annotators

Why not use BLEU, NIST or TER?
- small corpus
- dialogue utterances
- focus on sentence level
- only one reference translation

1) Preference rating (per utterance)

	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.

2) Annotation of differences between MT output and human gold standard and calculated edit distance

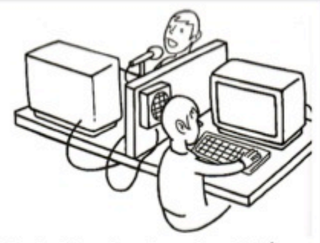
	GOOGLE			SYSTRAN		
add	110			111		
del	116			115		
edit	**226**			**226**		
	A1	A2	A3	A1	A2	A3
syn	25	23	23	27	20	27
adedit	**176**	**180**	**180**	**172**	**186**	**172**
wwo	3	3	3	4	4	4
ww	31	49	21	64	80	21
untr	10	11	11	0	0	0

Table 2: GOOGLE and SYSTRAN comparison including *edit* distance, *add*ed words, *del*eted words, adjusted edit distance (*adedit*), *syn*onyms, the number of wrong word orders (*wwo*), wrong words (*ww*) and untranslated words (*untr*) identified by annotators A1, A2 and A3.

Overall: the intrinsic evaluation shows a slight preference of the translations done by Google (cf. Kit and Wong 2008)

Extrinsic Evaluation

in a dialogue setting with the Wizard of OZ technique



Experimental Setting:

- WoZ prototyping tool (Schlögl et al. 10)
- 8 subjects talking to wizard via Skype
- wizard returns textual MT output
- data-logs of all interactions, questionnaires and expert interviews

1) Efficiency

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

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

2) Quality

	GOOGLE	SYSTRAN
Usage of the 'Original'-button:		
Uncomprehensible	2	3
Trust	3	4
Curiosity	5	8
Utterances that signal an interaction problem:		
'Sorry, but I have no information on that topic.'	1	1
'Sorry, I did not understand you. Could you say that again?'	5	2
'Do you want to go back a step?'	1	0
'Do you want to start over again?'	0	0

Table 4: Usage of the 'Original'-button and utterances that signal an interaction problem.

3) Task Success:
- all subjects completed the tasks
- measured through eight statements in the questionnaire
- no huge differences between the systems

4) Interviews:
- wrong word order recognized by all participants but not 'a big obstacle'
- 'Have a good day!' and 'Sorry' not translated by Google but not recognized
- 'Kilobyteantriebskraftgeschwindigkeit' unrecognized compound by Systran

Overall: subjects efficiently finished the task, no particular differences between Google and Systran

intrinsic evaluation is good to tune MT systems
but might not be meaningful to evaluate MT output in real live situations

This research is funded by Science Foundation Ireland (SFI) as part of the National Development Plan (NDP) 2007-2013