Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation

Douglas Jones, Paul Gatewood, Martha Herzog
MIT Lincoln Laboratory
HLT Research Group

Tamas Marius
Defense Language Institute
Foreign Language Center

1 Introduction

We present a new method for task-based speech-to-speech machine translation evaluation, in which tasks are defined and assessed according to independent published standards, both for the military tasks performed and for the foreign language skill levels used. We analyze task success rates and automatic MT evaluation scores for 220 role-play dialogs. Each role-play team consisted of one native English-speaking soldier role player, one native Pashto-speaking local national role player, and one Pashto/English interpreter. Machine translation (MT) and human translation (HT) conditions were assigned in a Latin Square design. Dialogs were assessed for language difficulty according to the Interagency Language Roundtable (ILR) speaking and listening skills.

The overall PASS score, averaged over all of the MT dialogs, was 44%. The average PASS rate for HT was 95%. Role players performed 20 tasks in 4 domains. The domain-level PASS scores ranged from 89% to 100% in the HT condition. For MT we observed 83% PASS rate in one of the four domains, Base Security (BS), with the remaining three domains ranging from 26% to 50% (Checkpoint Operations (CO), Civil Affairs (CA) and Situational Awareness (SA), an umbrella domain encompassing a variety of more complex tasks). The dialogs were human-scored in two main ways: (a) aggregate PASS/FAIL outcomes, and (b) a diagnostic assessment for specific communication initiatives. Inter-coder agreement for task PASS/FAIL scoring, which required an assessment of several performance measures per task, averaged 83%. Agreement for the specific communication initiatives was 98%. The PASS/FAIL scores for scenarios within the four domains are shown in Figure 1.

The dialogs were also assessed for language complexity. Scenarios with language complexity at the ILR Levels 1, 1+ and 2 had PASS scores of 94%, 100% and 92% respectively in the HT condition, as shown in Figure 2. For MT the overall results were 47%, 48% and 31%. In other words, MT performance is worse when the language is fundamentally more complex. The average BLEU score for English-to-Pashto MT was 0.1011; for Pashto-to-English it was 0.1505. BLEU scores varied widely across the dialogs. Scenario PASS/FAIL performance was also not uniform within each domain. Base Security scenarios did perform relatively well overall. On the other hand, although Civil Affairs scenarios did not perform that well on average, some of the scenarios were performed well with MT.

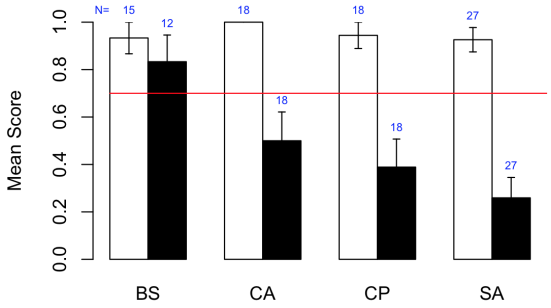

Figure 1: PASS/FAIL Results Across Domains

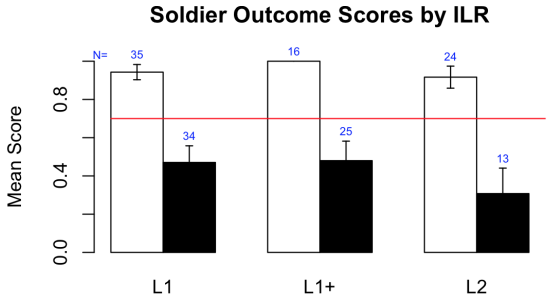

Figure 2: PASS/FAIL Scores by SME ILR Level for Speaking

Scenarios were of two general types: a basic definition without any complications, and a contrasting

This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

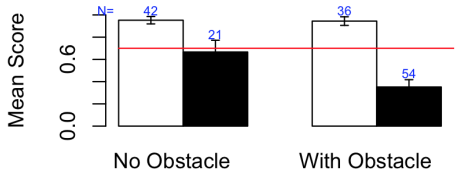

Figure 3: PASS/FAIL Results with Task Obstacles

definition with some type of obstacle, perhaps minor, that needed to be overcome in the communication. For example, in a basic Base Security scenario, a Local National may seek permission to pass a checkpoint with valid identification, but in a contrast scenario, he may lack the identification, but seek an alternative goal that does not involve passing the checkpoint. Overall PASS/FAIL results for the HT 95% for basic and 94% for contrast. For MT we observed 67% PASS for basic and 35% for contrast scenarios. The performance gap between HT at 94˜95% and MT with basic scenarios at 67% is 27% on average, whereas the difference between MT in basic scenarios and MT in contrasting scenarios is 32%, as shown in Figure 3.

The overall impression is that MT can work for some scenarios, but language complexity and task obstacles may drastically reduce performance, and in fact these factors can be as important as the contrast between machine translation and human translation. In other words, the effect of the scenario complexity is stronger than the effect of using MT instead of an interpreter. MT may provide needed communication when the only barrier for accomplishing a simple task is the language barrier. When the task has unexpected obstacles, MT is more likely to fail.

2 Design

We constructed a framework for evaluating speech-to-speech machine translation technology in a way that isolates spoken language used in performing standardized military tasks. Role players performed their duties via a push-to-talk communication system that routed human and machine-generated spoken language to the relevant role players and to a system monitor. Whether any given scenario used HT or MT was determined by the randomized position in the Latin Square design. The role players were as follows: (1) an English-speaking Subject Matter Expert (SME), a person who has experience with the various military tasks; (2) a Foreign Language Expert (FLE), a Pashto-speaking playing the role of the Local National; (3) an Interpreter (INT) who is able to provide immediate Pashto/English translation for the FLE and the SME. The SME and the FLE cannot hear each other; they communicate only via the interpreter providing human translation (HT) or machine translation (MT). Role players interacted via MT interface in all conditions in order to maintain consistency in communication behaviour. In the HT condition, the MT output was saved for further study, but not used in the role-play, the interpreter's live production of HT being swapped in.

The two human subjects engaged in a role-playing scenario. The SME took on the role of a US soldier charged to perform a certain duty based on relevant US Army training and doctrine. The SME spoke only in English. The FLE took on the role of an Afghan Local National. The FLE spoke only in Pashto. Both subjects were given common information about the scenario. Additional information was private to each role. Each subject also had a goal for the scenario. The job was for subjects to try to attain their goal through voice communication. Translation during the scenario was either machine-based or human-based.

2.1 Configuration

The FLE MT device was configured to prevent any English feedback to the role player. Although our role-players speak English in addition to Pashto, in real-world scenarios we certainly would not assume this capability. For this reason, the FLE was not allowed to monitor their English MT audio output. In this way the FLE role players, who were by necessity English speaking, could more accurately mimic a non-English speaking foreign national.

The diagram in Figure 4 shows the general construct of the communication setup.

The SME machine translation (MT) device was configured to provide textual feedback of both the automatic speech recognition (ASR) and back translation (English -> Pashto -> English). SME's were encouraged to make multiple attempts if necessary to achieve accurate ASR and if possible accurate back translation. Time constraints imposed a practical cap of 4-5 attempts per turn. On average, SMEs generated twice as many translations as were eventually used in communication, and FLEs generated 1.3 times as many.[1]

[1]The rejected "practice" MT output was not used in con-

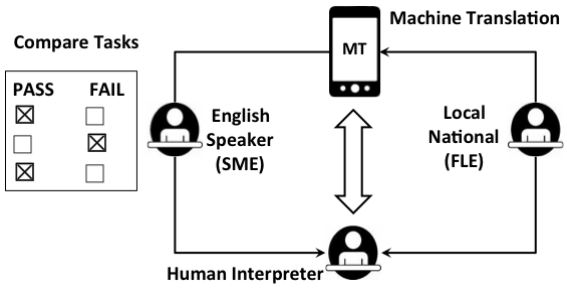

Figure 4: Evaluation Concept

The transcript in Table 1 shows examples of how role players used back translations to reject MT turns. Recall that the role players are able to see a back-translation into their native languages before sending the foreign language output to their counterpart for communication. If the back-translation looks too garbled, there is an opportunity to try again.[2] In Table 1, turns 3-5 shows how the SME rejected the first two translations based on the English back-translations (rejecting "Please help and take me to the explosion", trying again and also rejecting "Please help me gets to the explosion", and trying a third time, accepting "Please help me go to the explosion".

2.2 Terminology

By *task*, we mean US Army tasks as defined by TRADOC (Training and Doctrine) materials, such as Army Field Manuals, indexed by task identifier. The tasks used in our experiment are shown in Table 2. A *scenario* is a description of one of the these TRADOC tasks, specified in sufficient detail for the tasks to be performed consistently by role players. A *dialog* is the record of a specific role-playing event, a task performed by the role players. We refer to language expressed to initiate a communication goal as *communication initiatives*, regardless of whether it was understood or contributed to a passing score. For example, if the role player said, in English: "How many doctors are in the community?" the dialog would be scored positively for the initiative goal defined as "Does the role player ask about the local staff?" The overall *PASS/FAIL* goal for the dialog in this case was "Does the role player collect health information?" which required successful completion of several performance steps to receive a PASS. Inter-coder agreement for task PASS/FAIL scoring averaged 83%; agreement for the specific communication initiatives was 98%, as shown previously in Figure 1, the PASS/FAIL scores for scenarios.

2.3 Roles

The SME was expected to know the various training and doctrines necessary to perform the tasks within each scenario in a detailed and consistent fashion. They were expected to conduct these procedures as they would in a real world situation and maintain a respectful and courteous posture. In the case of scenarios where the SME and FLE goals are opposing, the SME tried to find a compromise or other way to assuage the FLE. SMEs used speech to communicate intent and instant messages (IM) to communicate physical context. The SME was shown how to operate the PC GUI including IM and the MT device during training.

The FLEs were expected to conduct themselves as they imagine they would under real world conditions while maintaining a respectful and courteous posture. If the FLE could not get a SME to agree to their goals as put forth in the scenario description the FLE would attempt to find a compromise or partial solution.

The interpreter provides real-time translation in both directions during HT scenarios and performs no interpretations during MT scenarios.

The monitor oversaw one set of role players (SME/FLE/INT). The Monitor assigned, began, ended and monitored all scenarios. They were responsible for reverting a turn if procedural mistakes were made. They also broadcast IMs for any physical action that adds to the shared context with the proper "SEE SME" or "SEE FLE" preface even when these were redundant to the player IMs. The monitor also took on the role of virtual SME and FLE by prompting the actual players via private IM any unaddressed communication goals before the end of the scenario. The monitor also scored the scenario. Runners oversaw individual role players and reported to the monitor. These were previous SMEs or FLEs with experience in performing the tasks. An experiment manager was responsible for the overall experiment. The manager typically monitored one of the role play teams in addition to guiding a second monitor when we had two parallel role

structing the reference translations and transcripts and hence was not used in calculating automatic scores, such as BLEU.

[2]For expository purposes in this paper, the English MT is shown in parenthesis for the FLE side, in addition to the Pashto back-translation seen by the role player. Recall that the FLE MT device was configured to only display Pashto, so for the FLE side, only the back-translations into Pashto were used.

Line	Role Player	Back-translation from MT Output Used by Role Players for Communication	"Practice" Back-translation MT Output not used by Role Players for Communication
1	SME:	I need your help	
2	*FLE:*	زه نه شم کولای *(I can not)*	
3	SME:		Please help and take me to the explosion
4	SME:		Please help me gets to the explosion
5	SME:	Please help me go to the explosion	
6	*FLE:*		ما ته محفوظ ځای وو خان دی *(I was khan is safe)*
7	*FLE:*	زه وېريږم *(I am afraid)*	
8	SME:		You will be okay come with me
9	SME:		You will not gephardt come with me
10	SME:	You will not be hurt come with me	
11	*FLE:*		صحیح ده *(Okay)*
12	*FLE:*	صحیح ده مننه *(Okay thank you)*	
13	SME:		Can you tell these people to stay far away
14	SME:		Tell all these people to stay back
15	SME:		Gets the people to back up
16	SME:		Tell the people to go away
17	SME:		Do you tell the people to go away
18	SME:	Tell the people to go away	
19	*FLE:*	ما سره مرسته وکړئ چې زه يې ته ولاړ شم *(Help me so that I can go)*	
20	SME:		Gets the people to back up
21	SME:		Move the people away
22	SME:	Move the people	
23	*FLE:*	پوه نه شوم *(I did not understand)*	
24	SME:		Help me move people away
25	SME:	Help me move people back	
26	*FLE:*	زه وېريږم تاسو سره مرسته نه شم *(I am afraid I can not help you)*	
27	SME:		Move away to safety
28	SME:		And then go to safety
29	SME:	Then move back so you do not get hurt	
30	*FLE:*	ما ته معلوم ځای وخت اما زه کومک نه شم کول *(I know where do I have time I did not help)*	

Table 1: Use of Back Translations Select MT Turns

Task ID	TRADOC Task		ILR Speaking Skill Level
		Civil Affairs	
CA2	331-38B-2020	Conduct a Local Medical Health Assessment	L2
CA4	331-38B-3015	Coordinate Handling of Supplies	L1+/2
CA5	331-38B-3033	Conduct Support to Civil Administration Operations	L2+
		Situational Awareness	
SA1	301-35M-1200	Implement Approach Strategies	L3
SA2	301-35M-1250	Assess Source for Truthfulness and Accuracy	L2
SA3	191-376-5126	Conduct Interviews	high L1
		Checkpoint Operations	
CP2	171-137-0001	Search Vehicles in a Tactical Environment	L1
CP6	191-376-5151	Control Access to a Military Installation	high L1
		Base Security	
BS1	191-376-4130	Operate a Roadblock as a Member of a Team	L1
BS4	191-376-5154	Respond to a Crisis Incident	L0+

Table 2: Task Inventory

playing teams.

2.4 Scenarios

The following lists show the information from the sample scenario that was provided to the role players. Both the SME and the FLE saw the material designated as "Shared Context".

- **SHARED_CONTEXT:** A US soldier is talking to a doctor. This Area of Operations (AO) is safe and secure.

The role players did not see each other's knowledge and goals. The SME saw these descriptions:

- **SME_KNOWS:** You are a Civil Affairs Solider assigned to a civil-military operations center. You know how to conduct a Local Medical Health Assessment. Today you are concentrating on collecting human health information only.
- **SME_GOAL:** Follow the procedures for Conducting a Local Human Health Assessment. Collect health information including local facility names, the number of health workers and the nearest pharmacy. Determine the size of the population. Identify any endemic diseases and the leading causes of death.

Likewise, the FLE saw only this relevant part of the scenario description:

- **FLE_KNOWS:** You are the only doctor here with responsibility for the two villages and the surrounding area. Your office is the only clinic and pharmacy. You have a meager stock of only the most basic medications. There are about 200 extended families. Twenty percent of the population is over 65 years old, or about 250 people. Children under the age of 12 number about 400. Cholera is the leading cause of death. There are scattered cases of hepatitis B. You have also seen a rise in cases of brucellosis.
- **FLE_GOAL:** Describe the medical situation of the population in your area.

Neither the SME nor the FLE see the scoring criteria during the role-play in order to avoid overscripting the dialogs. A full day of training was provided to the SME to cover the requirements according to the standard definitions of task, conditions, standards, performance steps and performance measures. After the role-play, the SME was assigned PASS/FAIL scores for the specific goals shown in Figure 5. We defined goals for the FLE to be scored in a similar fashion, although these are obviously not part of the training materials for soldiers. The FLE goals for this sample scenario are also shown in Figure 5.

	Sample Soldier SME Goals
1	Collect health information.
2	Ask about the local facilities.
3	Ask about the local staff.
4	Ask about the nearest pharmacy.
5	Ask about the size of the population.
6	Ask about any endemic diseases.
7	Ask about the leading cause of death.

	Sample FLE Goals
1	Describe the local facilities.
2	Describe the local staff.
3	Describe the nearest pharmacy.
4	Convey data on the population.
5	Describe any endemic diseases.
6	Describe the leading cause of death.
7	Convey health information.

Figure 5: SME Goals for Sample Scenario

Table 3 shows the list of scenarios in the experiment.[3]

2.5 Sample Transcripts

Table 4 shows three sample transcripts. The first one shows a transcript of the human interpreter for a Base Security scenario. The second one is fairly fluent machine translation in a Base Security scenario. This is the same dialog shown in Table 1 which showed how some MT output is rejected using back translations. The third shows a lower quality MT interaction which causes some difficulty for the role players in the Civil Affairs scenario.

2.6 Scoring

Dialogs were scored in three different ways for each side of the conversation. *Communication Initiatives* are scored with one side of the conversation; they are scored as PASS if the Role Player attempts to communicate a particular objective, regardless of whether the FLE understands it or does anything in response. For example: "Does the SME ask about the local facilities?". *World Goals* are scored with

[3]Three of the twenty main scenarios were removed from both MT and HT scores because the PASS/FAIL goals were not met in

ScenID	Scenario Title and *Domain*
	Base Security
BS11	Slow US Convoy Blocking Road
BS45	Respond to a Crisis Incident
BS46	Local National wants to help in crisis
	Civil Affairs
CA21	Human Health Assessment
CA22	Animal Health Assessment
CA51	Return Displaced Civilians
CA54	Discuss Situation regarding Rule of Law
	Checkpoint Operations
CP21	Family Vehicle at Checkpoint
CP22	National with Borrowed Vehicle
CP61	Local National Wants to See Commander
CP66	Local National Wants Job
	Situational Awareness
SA11	Approach Tea Shop Owner for Information
SA12	Speak with Villager for Information
SA21	Assess Villager as Source for Information
SA22	Assess Insurgent Cell Member as Source
SA12	Interview Escaped Detainee
SA13	Local National Police Training

Table 3: Scenario Inventory

respect to both side of the conversation for one role player; they scored as a PASS if the role player's real world outcome was met. The world goal may require action or communication on the part of the both role players in order to be assigned a PASS. They require successful two-way communication to meet the goal. For example: "Does the SME collect health information?". *PASS/FAIL Outcomes* is aggregate score that is scored scored as PASS only if ALL of the role player's world goals were met, otherwise scored as FAIL. It requires successful two-way communication over the dialog as a whole.

The abbreviations for the score types are as follows: **SC**: SME Communication Initiative; **SW**: SME World Goal; **SO**: SME PASS/FAIL Outcome.

Scores for Soldier Communication Initiatives and PASS/FAIL outcomes for all of the tasks are shown in Figure 6 and Figure 7 respectively. The key observations are: (1) the soldier role players generally succeed in their Communication Initiatives, both for MT and HT. However, the overall outcomes vary greatly on a scenario by scenario basis. The clearest distinction is between the Base Security (BS) scenarios at the more successful end, compared with the more complex Situational Awareness (SA) tasks with lower success rates.

	Human Interpreter
	Civil Affairs Scenario CA12, Team 6
SME:	how many men and trucks can you provide ?
FLE:	*two cars four people four guns .*
SME:	do you know the roads very well ?
FLE:	*yes I know the roads very well .*
SME:	can you make sure nobody steals the supplies ?
FLE:	*okay . I'm trying my best .*
SME:	do you know the safest way to get there ?
FLE:	*yes I know a safe way .*
SME:	how many men can you provide ?
FLE:	*I can provide you with four men .*
SME:	how many trucks ?
FLE:	*two trucks .*
SME:	sounds good . I like the plan .
FLE:	*no problem . I'm thankful to you .*
	Better Machine Translation
	Base Security Scenario BS45, Team 6
SME:	I need your help .
FLE:	*I can not*
SME:	please help me go to the explosion .
FLE:	*I am afraid*
SME:	you will not be hurt . come with me .
FLE:	*Okay thank you*
SME:	tell the people to go away .
FLE:	*Help me so that I can go*
SME:	move the people .
FLE:	*I did not understand*
SME:	help me move people back .
FLE:	*I am afraid I can not help you*
SME:	then move back so you do not get hurt .
FLE:	*I know where do I have time I did not help*
	More Difficult Machine Translation
	Civil Affairs Scenario CA42, Team 1
SME:	I send medical equipment and food .
FLE:	*You have old system*
SME:	do you have any suggestions ?
FLE:	*And I have done*
SME:	are you familiar with roads in this area ?
FLE:	*And this area is all dirt paved roads we know*
SME:	do you know a good road ?
FLE:	*And I know all this garbage paved roads*
SME:	where are the criminal groups near these roads ?
FLE:	*The roads fees criminal vehicles there*
SME:	what can you give me for security ?
FLE:	*I have four guys about guns trucks for you*
SME:	I will use these gun trucks .
FLE:	*That's very good*
SME:	how many trucks and how many men do you have ?
FLE:	*How many times the trucks and four guns person*
SME:	how many trucks ?
FLE:	*Four men and two ways*
SME:	how many trucks only ?
FLE:	*There are four two trucks*
SME:	only two vehicles ?
FLE:	*There are four guard with them*
SME:	I understand your plan and will take your supplies .
FLE:	*That's very good*

Table 4: Sample Transcripts

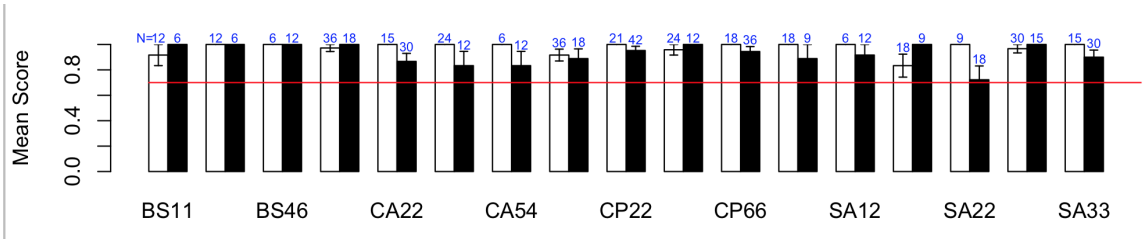

Figure 6: Communication Initiative Scores for Scenario

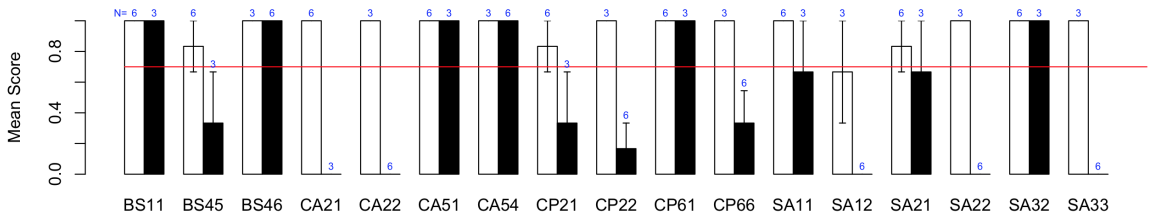

Figure 7: PASS/FAIL Scores for Scenarios

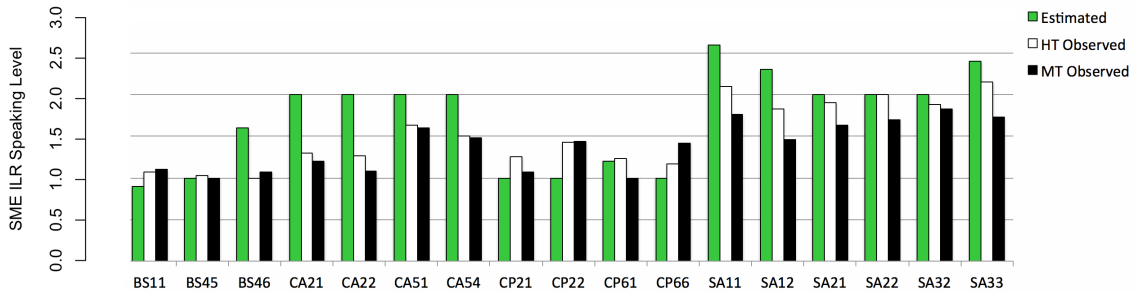

Figure 8: ILR Speaking Skills Estimated and Observed for Scenarios

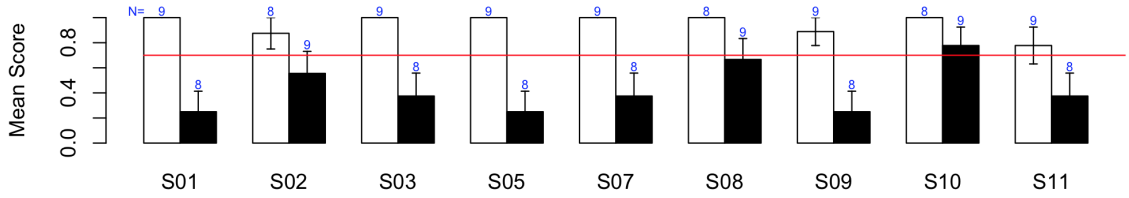

Figure 9: PASS/FAIL Scores by Subject Team

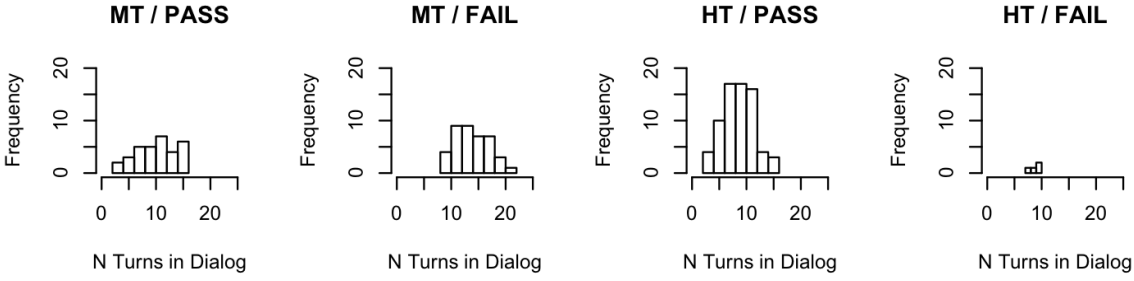

Figure 10: Length of Dialogs and PASS/FAIL Scores

2.7 Data Triage

There were three steps of data triage. First, we dropped two of the original twelve SME teams whose overall PASS average in the HT condition was lower than 70%. In order to assess the relative effect of machine translation, we required that the role players be able to perform the tasks required. In this triage step, we dropped all role plays for those two teams, both MT and HT. Second, we also dropped the subgoals for which the average success rate was less than 70%: of the 209 goals, 20 were dropped. Third, the second step of subgoal triage meant that 3 of the 20 scenarios lacked an overall PASS/FAIL score, so these were also excluded. Future experiments would repair the subgoals and associated overall PASS/FAIL scoring, and should only need the first triage step to exclude role-players who could not complete the tasks in the HT condition.

2.8 ILR Assessment of TRADOC Tasks

The tasks were assessed according to ILR skill level likely needed for successful task completion, as shown in previously in Table 2. The Speaking and Listening skill estimates were generally very close. The actual Speaking ILR levels observed for both HT and MT are shown in Figure 8. In general, the language used for MT was slightly less complex than for HT, and the levels were somewhat lower than what was estimated in advance of the experiment.

As might be expected, the number of words used in communication correlates somewhat with the ILR Speaking levels, as shown in Figure 11, with $R^2 = 24\%$. In other words, the role player speaks more when the required language skill is more challenging.

2.9 Dialog Length

In Figure 10, a histogram of the number of turns needed to complete the dialog is shown for four cases. First, the number of turns used in dialogs receiving a PASS score. These are usually completed in under 15 turns. Failing dialogs never finished early; they required as many as 20 or more turns until they reached the time limit for the role-play, at which point they moved on. Passing dialogs in the HT condition had about as many turns as the passing dialogs in the MT condition. Very few HT dialogs failed; these were due to missed subgoals on the part of the role-players, rather than a general failure to communicate.

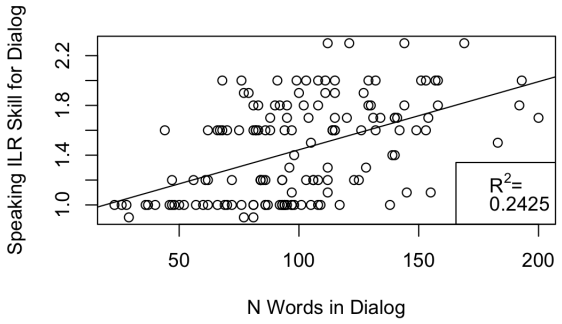

Figure 11: Words in SME Transcript and SME ILR Level for Speaking

2.10 Subject Variation

We observed a noticeable amount of variation by subject, as shown in Figure 9. The highest performing team was S10 at 78% PASS rate for MT. A binomial test, using the overall PASS rate of 44% for MT, shows that this result would be achieved by chance just under 5% of the time, not quite enough to expect that this team is doing something special.

2.11 Automatic Scores

Machine translation is most typically evaluated with automatic methods such as BLEU and METEOR. They are typically used because of convenience and interpretability within the research and development community, and we have performed those measurements as well, as shown in Figure 12.[4]

The automatic scores were not correlated with the overall PASS/FAIL scores, either for BLEU or for METEOR. However, BLEU and METEOR scores correlate with each other, more strongly when compared at the dialog level and less strongly when the dialog scores are averaged over the role-player teams per scenario. One fundamental issue is one of granularity. The PASS/FAIL scores are assigned for each dialog, a relatively large unit. These are averaged over the performance by several role players, half of whom perform the scenario in the MT condition. Automatic scores such as BLEU and METEOR can be averaged over many different levels of granularity. However, collapsing the automatic scores to the scenario level throws away information. As might be expected, the BLEU scores correlate METEOR better at the dialog transcript level ($R^2 = 66\%$) than at the scenario level, which av-

[4]We did not try to normalize the transcripts in producing these scores.

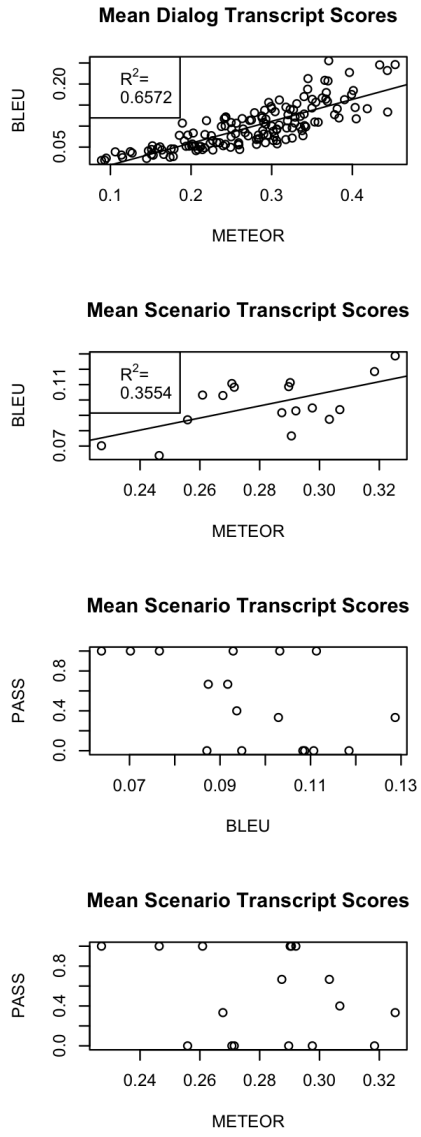

Figure 12: PASS/FAIL compared with BLEU and METEOR

erages the dialog performance over the role players for that scenario ($R^2 = 36\%$). Fitting a linear regression line onto the BLEU and METEOR data on this data set would show a negative correlation with both BLEU and METEOR. However, due to the small number of scenarios (there are only 20 scenarios in the role play), it would be a mistake to read too much into a negative correlation. The appropriate observation is to note that the automatic scores are not a reliable predictor of PASS/FAIL scores.

It is plausible, however, that a negative correlation could exist. One other fundamental issue is the difference between interactive machine translation and batch processing. In these interactive scenarios, if a role player has the opportunity to try again after recognizing a bad translation. In fact, the more the role-player works to repair the communication, the longer trail that might be left of failed translations. These failures contribute to the BLEU score and may outweigh the successful translations in their quantity. For example, in the transcript in Table 5, the SME is very persistent in working around communication failures. Ultimately the SME was able to achieve this goal, despite failed interchanges within the dialog. Each of those failed interchanges would have contributed to a worse BLEU score despite an overall PASS for that dialog.

2.12 Contrast with Other Manual Methods

Manual methods such as Concept Transfer Rate have been used for speech-to-speech machine translation evaluation, for example Sanders et al. (2011). Concept transfer rates are based on counting the number of key concepts communicated within a fixed time period. The new method that we describe here departs from these conventions by allowing the role players greater freedom to accomplish their tasks. One of the primary motivations was to avoid the risk of over-scripting role player behavior. For example, if communication breaks down early in the scenario, we do not want the role players to work from a script to artificially repair the situation. The main risk to the experiment is if the role players fail to properly perform their tasks even in the HT condition. To mitigate this risk, we only employed role players who had performed the required tasks in a recent military deployment, and we provided a full day of training using 10 additional scenarios from the same domains as the evaluation scenarios.

3 Acknowledgements

This work builds on prior work on ILR-Based Machine Translation evaluation performed by MIT Lincoln Laboratory, the National Institute of Standards and Technology and the US Army Machine Foreign Language Translation System program. We used the "Walkie-Talkie Over IP" (WOIP) system developed by Wade Shen as the communication system, and was instrumental in the design of the previous experiments for MFLTS, for which we wish to thank him. The original scenarios were developed at a military training exercise in 2009 with assistance from Patrick O'Malley, a retired US Army officer who supports TRADOC at Fort Huachuca, and Neil Granoien, a retired Vice Chancellor from the DLI who was working with us at that time. The

	Civil Affairs Scenario CA54, Team 9		
SME:	thank you for meeting with me .	SME:	by having your public trust your court your town will be more functional .
FLE:	*Not coming on us*	*FLE:*	*We have*
SME:	what did you say ?	SME:	what do you need from us to help you ?
FLE:	*All the last*	*FLE:*	*In our security is*
SME:	what's wrong with your courts ?	SME:	is all you need security ?
FLE:	*Every work is security*	*FLE:*	*We have*
SME:	why are trials not happening ?	SME:	what did you say ?
FLE:	*They are hidden nervous*	*FLE:*	*The security is for the country*
SME:	why are you nervous ?	SME:	I'm concerned about your court system not your country .
FLE:	*Put under the work*	*FLE:*	*If we have security*
SME:	your public needs to trust you and be safe .	SME:	to confirm all you need is security from me .
FLE:	*He can be*	*FLE:*	*We were*
SME:	what did you say ?	SME:	okay . we will send people over tomorrow .
FLE:	*Our security officials work*	*FLE:*	*Thank you very much it*
SME:	will you let us help you ?		
FLE:	*It is*		

Table 5: Overcoming Garbled Translation

role plays were conducted in Professor Ted Gibson's laboratory at MIT. We also wish to thank the role players: the military veterans who participated as SMEs, and the native Pashto speakers who participated as FLEs.

4 Conclusion

The key lessons we learned in the experiment is that it is not enough for the role players to express themselves. What they say has to be understood by their conversational partner for the scenario to be completed successfully. Moreover, success depended as much on task simplicity as it did upon the translation condition, given that 67% of the base case scenarios were successfully completed using MT, but only 35% of the contrasting scenarios with even minor obstacles received passing scores. In other words, we observed that MT had the greatest chance of success when the task was simple and the language complexity needs were low.

We feel that our earlier work suggested that ILR proficiency Level 2 was the ideal level for text translation of existing texts. That may be because concrete facts are more readily transferable from one language to another. Similarly, in speech to speech translation, Level 1 is may be ideal because such language focuses on the everyday, the here and now, the immediate situation discernible to all parties. The task is to work out a solution to a simple issue or problem within this situation. (e.g. *The road is blocked; how do I get home?*)

The implication of the performance variation shown in these results is that the technology should be tested in advance for specific situations in which it might be used, and not to assume that it just "works" for a particular broad domain.

5 Bibliography

1. Jones, Douglas, et al. (2007) ILR-based MT Comprehension Test with Multi-Level Questions. NAACL 2007.

2. Sanders, Gregory, et al. (2011) Evaluation Methodology and Metrics Employed to Assess the TRANSTAC two-way speech-to-speech translation systems. Computer Speech and Language.

3. Papineni, Kishore, et al. (2001) "Bleu: a Method for Automatic Evaluation of Machine Translation" IBM Computer Science Research Report RC22176 (W0109-022) 9/ 17/2001

4. US Army Doctrine and Training Publications, http://armypubs.army.mil/doctrine/