ILR-Based MT Comprehension Test with Multi-Level Questions

Douglas Jones, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen,
Edward Gibson and Michael Emonts

MIT Lincoln Laboratory
Lexington, MA 02420
{DAJ,Arvind,SWade}@LL.MIT.EDU
MHerzog2005@comcast.net

DLI Foreign Language Center
Monterey, CA 93944
{Hussny.Ibrahim,Michael.Emonts}
@monterey.army.mil

MIT Brain and Cognitive
Sciences Department
Cambridge MA, 02139
EGibson@MIT.EDU

Abstract

We present results from a new Interagency Language Roundtable (ILR) based comprehension test. This new test design presents questions at multiple ILR difficulty levels within each document. We incorporated Arabic machine translation (MT) output from three independent research sites, arbitrarily merging these materials into one MT condition. We contrast the MT condition, for both text and audio data types, with high quality human reference Gold Standard (GS) translations. Overall, subjects achieved 95% comprehension for GS and 74% for MT, across 4 genres and 3 difficulty levels. Surprisingly, comprehension rates do not correlate highly with translation error rates, suggesting that we are measuring an additional dimension of MT quality. We observed that it takes 15% more time overall to read MT than GS.

1 Introduction

The official Defense Language Proficiency Test (DLPT) is constructed according to rigorous and well-established principles that have been developed to measure the foreign language proficiency of human language learners in U.S. Department of Defense settings. In 2004, a variant of that test type was constructed, following the general DLPT design principles, but modified to measure the quality of machine translation. This test, known as the DLPTstar (Jones et al, 2005), was based on authentic Arabic materials at ILR text difficulty levels 1, 2, and 3, accompanied by constructed-response questions at matching levels. The ILR level descriptors, used throughout the U.S. government, can be found at the website cited in the list of references. The text documents were presented in two conditions in English translation: (1) professionally translated into English, and (2) machine translated with state-of-the art MT systems, often quite garbled. Results showed that native readers of English could generally pass the Levels 1 and 2 questions on the test, but not those at Level 3. Also, Level 1 comprehension was less than expected, given the low level of the original material. It was not known whether the weak Level 1 performance was due to systematic deficits in MT performance at Level 1, or whether the materials were simply mismatched to the MT capabilities.

In this paper, we present a new variant of the test, using materials specifically created to test the capabilities of the MT systems. To guarantee that the MT systems were up to the task of processing the documents, we used the DARPA GALE 2006 evaluation data sets, against which several research sites were testing MT algorithms. We arbitrarily merged the MT output from three sites. The ILR difficulty of the documents ranged from Level 2 to Level 3, but the test did not contain any true Level 1 documents. To compensate for this lack, we constructed questions about Level 1 elements (e.g., personal and place names) in Level 2 and 3 documents. A standard DLPT would have more variation at Level 1.

2 Related and Previous Work

Earlier work in MT evaluation incorporated an informativeness measure, based on comprehension test answers, in addition to fluency, a measure of output readability without reference to a gold standard, and adequacy, a measure of accuracy with reference to a gold standard translation (White and O'Connell, 1994). Later MT evaluation found fluency and adequacy to correlate well enough with automatic measures (BLEU), and since comprehension tests are relatively more expensive to create, the informativeness test was not used in later

Proceedings of NAACL HLT 2007, Companion Volume, pages 77–80,
Rochester, NY, April 2007.

MT evaluations, such as the ones performed by NIST from 2001-2006. In other work, task-based evaluation has been used for MT evaluation (Voss and Tate, 2006), which measures human performance on exhaustively extracting 'who', 'when', and 'where' type elements in MT output. The DLPT-star also uses this type of factual question, particularly for Level 2 documents, but not exhaustively. Instead, the test focuses on text elements most characteristic of the levels as defined in the ILR scale. At Level 3, for example, questions may concern abstract concepts or hypotheses found in the documents. Applying the ILR construct provides Defense Department decision makers with test scores that are readily interpretable.

3 Test Construction and Administration

In this paper, we present a new test, based entirely on the DARPA GALE 2006 evaluation data, selecting approximately half of the material for our test. We selected twenty-four test documents, with balanced coverage across four genres: newswire, newsgroups, broadcast news and talk radio. Our target was to have at least 2500 words for each genre, which we exceeded slightly with approximately 12,200 words in total for the test. We began with a random selection of documents and adjusted it for better topic coverage. We constructed an exhaustive set of questions for each document, approximately 200 questions in total. The questions ranged in ILR difficulty, from "0+, 1,1+, 2, 2+ and 3, with Levels 0+, 1 and 1+ combined to a pseudo-level we called L1~, providing four levels of difficulty to be measured. We divided the questions into two sets, and each individual subject answered questions for one of the sets. The test itself was constructed by a DLPT testing expert and a senior native-speaking Arabic language instructor, using only the original Arabic documents and the Gold Standard translations. They had no access to any machine translation output during the test construction or scoring.

In August 2006, we administered the test at MIT to 49 test subjects who responded to announcements for paid experimental subjects. The subjects read the documents in a Latin square design, meaning that each subject saw each document, but only in one of the two conditions, randomly assigned. Subjects were allowed 5 hours to complete the test. Since the questions were divided into two sets for each document, the actual set of 49 subjects yielded approximately 25 "virtual subjects" reading the full list of 228 questions. The mean time spent on testing, not counting breaks or subject orientation, was 2.5 hours; fastest was 1.1 hours, slowest was 3.4 hours.

The subject responses were hand-graded by the two testing experts, following the pre-established answers in the test protocol. There was no pre-assessment of whether information was preserved or garbled in the MT when designing questions or responses in the test protocol. The testing experts were provided the reference translations and the original Arabic documents, but not the MT during scoring. Moreover, test conditions were masked in order to provide a blind assessment. The two testing experts provided both preliminary and final scores; multiple passes provided an opportunity to clarify the correct answers and to normalize scoring. The scoring agreement rate was 96% for the final scores.

4 Overall Results

The overall result for comprehension accuracy was 95% for subjects reading the Gold Standard translation and 74% for reading Machine Translation, across each of the genres and difficulty levels. The comprehension accuracy for each genre is shown in Figure 1. The two text genres score better than the audio genres, which is to be expected because the audio MT condition has more opportunities for error. Within each modality, the more standard, more structured genre fares better: newswire results are better than newsgroup results, and the more structured genre of broadcast news scores better than the less constrained, less structured conversations present in the talk radio shows.

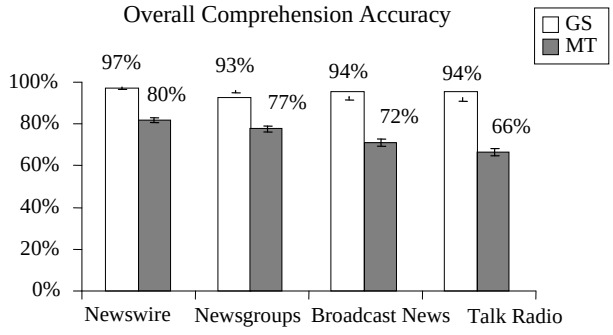

Figure 1. Comprehension Accuracy per Genre

The break-down by ILR level of difficulty for each question is shown in Figure 2. The general trend is consistent with what has been observed previously (Jones et al. 2005). The best results are at Level 2; Level 1 does well but not as well as expected. Thus the test has provided a key finding, which is that MT systems perform more poorly on Level 1, even when the data is matched to their capabilities. Level 3 is very challenging for the MT condition, and also more difficult in the GS condition. Using a standard 70 percent passing threshold, responses to questions on all MT documents, except for Level 3, received a passing grade.

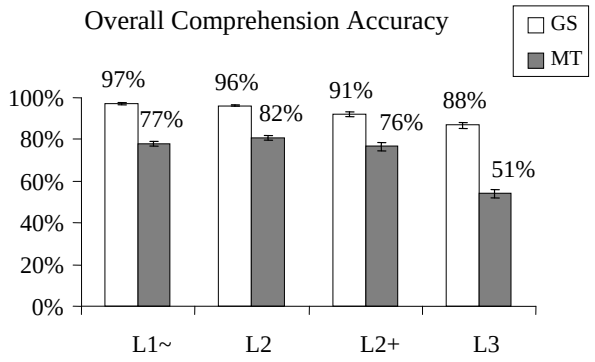

Figure 2. Comprehension Accuracy per Level.

To provide a snapshot of the ILR levels: L1 indicates sentence-level comprehensibility, and may include factual local announcements, etc.; L2 indicates paragraph-level comprehensibility; factual/ concrete, covering a wide spectrum of topics (politics, economy, society, culture, security, science); L3 involves extended discourse comprehensibility; the ability to understand hypotheses, supported opinion, implications, and abstract linguistic formulations, etc.

It was not possible to balance Level 3 documents across genres within the GALE evaluation data; except for those taken from Talk Radio, most documents did not reach that level of complexity. Hence, genre and difficulty level were not completely independent in this test.

5 Comprehension and Translation Error

We expect to see a relationship between comprehension rates and translation error. In an idealized case, we may expect a precise inverse correlation. We then compared comprehension rates with Human Translation Error Rate (HTER), an error measure for machine translation that counts the number of human edits required to change system MT output so that it contains all and only the information present in a Gold Standard reference (NIST, 2006). The linear regression line in Figure 3 shows the kind of inverse correlation we might expect. Subjects lose about 12% in comprehension for every 10% of translation error. The R^2 value is 33%. The low correlation suggests that the comprehension results are measuring a somewhat independent aspect of MT quality, which we feel is important. HTER does not directly address the facts that not all MT errors are equally important and that the texts contain inherent redundancy that the readers use to answer the questions. For exploratory purposes, we divide the graph of Figure 3 into four quadrants. Quadrant I and IV contain expected behavior: 122 data points of good translations and good comprehension results versus 43 points of bad translations and poor comprehension. Q-II has 24 robust points: the translations have high error, but somehow managed to contain enough well-translated words that people can answer the questions. Q-III has 28 fragile points: the few translation errors impaired comprehension.

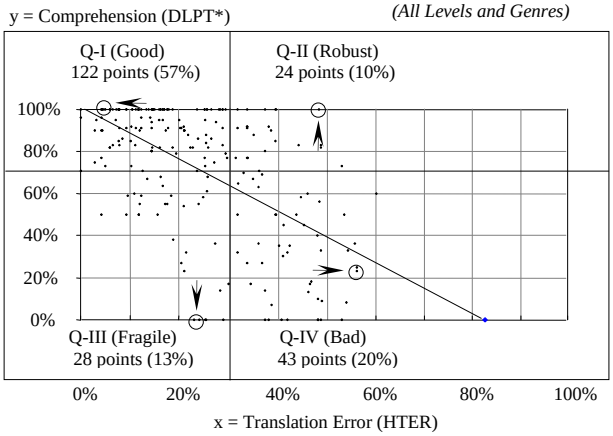

Figure 3. Comprehension vs. Translation Error.

We point out that there is a 1-to-1 mapping between comprehension questions and individual sub-passages of the documents in the data. Each point in Figure 3 plots the HTER of a single segment versus the average comprehension score on the corresponding question. The good and bad items are essentially a sanity-check on the experimental design. We expect to see good comprehension when translations are good, and we expect to see poor comprehension when translations are bad. Next we will examine the two other types: fragile and robust translations.

A fragile translation is one that has a good HTER score but a bad comprehension score. A sample fragile translation is one from a broadcast news which asks for a particular name: the HTER was a respectable 24%, but the MT comprehension accuracy was a flat 0%, since the name was missing. Everyone reading GS answered correctly.

A robust translation is one that has a bad HTER score but still manages to get a good comprehension score. A sample robust translation is one drawn from a posting providing instructions for foot massage. The text was quite garbled, with an HTER score of 48%, but the MT comprehension accuracy was a perfect 100%. Everyone reading the GS condition also answered the question correctly, which was that one should start a foot massage with oil. We note in passing that the highest error rate for a question with 100% comprehension is about 50%, shown with the up-arrow in Figure 3. We should be surprised to see any items with 100% comprehension for HTER rates above 50%, considering Shannon's estimate that written English is about 50% redundant. We expect that MT readers are making use of their general world knowledge to interpret the garbled MT output. A challenge is to identify robust translations, which are useful despite their high translation error rate.

6 Detailed Discussion

In this section we will discuss several aspects of the test in more detail: the scoring methodology, including a discussion of partial credit and inter-rater agreement; timing information; questions about personal names.

Each correct answer was assigned a score of 1, and each incorrect answer was assigned a score of 0. Partial credit was assigned on an ad-hoc basis, but normalized for scoring by assigning all non-integer scores to 0.5. This method yielded scores that were generally at the midpoint between binary scoring, in which non-integer scored were uniformly mapped either harshly to 0 or leniently to 1, the average difference between harsh and lenient scoring being approximately 11%. Inter-rater agreement was 96%.

The testing infrastructure we used recorded the amount of time spent on each document. The general trend is that people spend longer on MT than on GS. The mean percentage of time spent on MT compared with GS is 115% per item, meaning that it takes 15% more time to read MT than GS. The standard error was 4%. The median is 111%; minimum is 89% and maximum is 159%. In future analysis and experimentation we will conduct more fine-grained temporal estimates.

As we have seen in previous experiments, the performance for personal names is lower than for non-names. We observed that the name questions have 71% comprehension accuracy, compared with the 83% for questions about things other than personal names.

7 Conclusions and Future Work

We have long felt that Level 2 is the natural and successful level for machine translation. The ability to present concrete factual information that can be retrieved by the reader, without requirements for understanding the style, tone, or organizational pattern used by the writer seemed to be present in the previous work. It is worth pointing out that though we have many Level 1 questions, we are still not really testing Level 1 because the test does not contain true Level 1 documents. In future tests we wish to include Level 1 documents and questions.

Continuing along these lines, we are currently creating two new tests. We are constructing a new Arabic DLPT-star test, tailoring the document selection more specifically for comprehension testing and ensuring texts and tasks are at the intended ILR levels. We are also constructing a Mandarin Chinese test with similar design specifications. We intend for both of these tests to be available for a public machine translation evaluation to be conducted in 2007.

References

Doddington, G. 2002. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. Proceedings of HLT 2002.

NIST 2006. *GALE Go/No-Go Eval Plan*; www.nist.gov/speech/tests/gale/2006/doc/GALE06_evalplan.v2.pdf

Jones, D. A., W. Shen, et al. 2005a. *Measuring Translation Quality by Testing English Speakers with a New DLPT for Arabic.* Int'l Conf. on Intel. Analysis.

Interagency Language Roundtable Website. 2005. *ILR Skill Level Descriptions*: http://www.govtilr.org

Voss, Clare and Calandra Tate. 2006. *Task-based Evaluation of MT Engines*. European Association for Machine Translation conference.

White, JS and TA O'Connell. 1994. *Evaluation in the ARPA machine translation program: 1993 methodology*. Proceedings of the HLT workshop.