USER-CENTERED EVALUATION FOR MACHINE TRANSLATION OF SPOKEN LANGUAGE

David D. Palmer

Virage Advanced Technology Group
300 Unicorn Park
Woburn, MA 01801
dpalmer@virage.com

ABSTRACT

In this paper we discuss a user-centered method for qualitative comparison of machine translation systems based on rankings of system output. System ranking requires no reference transcript, which can be expensive to generate and difficult to define for spoken language input. Ranking can be performed by monolingual users with no training in machine translation evaluation. We present results of experiments ranking four Arabic-to-English and three Mandarin-to-English machine translation systems processing spoken language transcripts with word error rates of 20-30%.

1. INTRODUCTION

The quality of machine translation (MT) system output has improved dramatically in recent years, and this improvement has enabled the use of machine translation output in real-world applications, such as document filtering, information extraction, and summarization. Improvements in the speed of MT systems have resulted in the integration of MT into real-time applications processing live text and audio feeds. With this rapid improvement in both speed and quality has come a dramatic increase in the number of people working with machine-generated language data on a regular basis. As with the migration of automatic speech recognition (ASR) systems from the lab to the real world over the past decade, widespread use of MT data requires an understanding of user perceptions of MT quality. There has been a great deal of research in MT evaluation, but high scores on a formal evaluation do not guarantee that "novice" users will find system output useful for their tasks. In this paper we present preliminary user-centered experiments with several different MT systems, with the goal of determining which characteristics of MT output improve user perception of the quality of MT systems.

2. MACHINE TRANSLATION EVALUATION

The evaluation of machine translation systems is historically expensive and time-consuming, requiring human reference translations and human scoring of translation quality. MT evaluations have focused on two particular aspects of the system output: fluency and adequacy [6]. Fluency is the naturalness of the output in the target language; adequacy is the extent to which the target language output contains the information in the source language input. Since both fluency and adequacy are subjective measures of quality, the evaluation must be carried out by expert humans trained to assign meaningful scores based on linguistic criteria. In some cases the humans are bilingual and are able to compare the source and target texts directly; in other cases a monolingual human compares the MT output to a set of reference translations produced by bilingual human experts.

The BLEU method [5] provided the first automated approach to evaluation by comparing n-grams in MT output to those in a set of human reference translation. This enabled rapid system development and evaluation without the need for human scoring of output quality. However, while BLEU provides a means for scoring a particular system automatically, the evaluation is still dependent on the existence of a set of human translations for comparison.

The vast majority of formal MT evaluations have been carried out with "clean" text data, passages generated and edited by humans, with consistent capitalization and punctuation. There has been some work in evaluating MT for spoken language data [1,4], in which the input to the MT system is the transcript generated by an automatic speech recognition system. This work applies clean data evaluation techniques to the noisy data problem. However, noisy data input complicates the traditional MT evaluation paradigm, since adequacy is ill-defined: is a good translation one that faithfully reproduces errors in the ASR transcript or one that recreates the lost information from the original spoken audio? Consider a case where the spoken word "Iraq" is output by the ASR system as "a rock." It is very

0-7803-8874-7/05/$20.00 ©2005 IEEE

unlikely that any MT system would generate a target language passage containing "Iraq" given "a rock" as input, and evaluating the system using the traditional MT paradigm would not be a useful measure of the MT system performance. In this work we seek to develop a new evaluation paradigm for spoken language MT that can take these factors into account.

3. USER-CENTERED EVALUATION

Our work in MT evaluation is carried out in the context of the Enhanced Video Text and Audio Processing (eViTAP) project [3], a fully-automated real-time multilingual broadcast news processing system. The eViTAP system combines speech recognition, machine translation, and cross-lingual information retrieval components to enable real-time navigation of live English, Arabic, and Mandarin news sources. System users can retrieve news stories in any of the source languages, play the corresponding video, and view synchronized transcripts in both the broadcast source language (via ASR) and in English (via MT of the ASR output). The readability of the MT output is thus a key factor in the system usability and the ability of the user to identify relevant news stories.

The eViTAP system is designed to work seamlessly with several different ASR systems and MT systems processing the audio from the live news broadcasts. System users are typically monolingual English speakers and therefore could provide no information regarding the quality of the Arabic or Mandarin ASR output or the adequacy of the MT output in conveying the information in the spoken language. Since system users depend on the language transcripts to identify relevant news stories and to prepare English-language reports, our goal is to assist the users in determining which MT system would be most likely to provide the best transcript in the context of the eViTAP data flow.

Our user-centered evaluation focused on determining relative MT system rankings for a set of output passages. The evaluators were English-speaking users of the eViTAP system, with no knowledge of Arabic or Mandarin and no training in formal evaluation of MT performance. Because the evaluation data was "fresh" data produced by the ASR+MT cascade, there was no reference transcript for either the source or the target language. However, one key element was that the users had access to all MT outputs for a single input and could compare the passages directly and determine how easily they could get the information from each passage. Although it was impossible to determine which output passage had the most information from the source passage, it was possible to determine, with some confidence, which output was missing information that the others had.

4. EXPERIMENTS

In our experiments, each English-speaking user was presented with the output from different MT systems for the same input passage of ASR transcription. No reference translation was provided, but the user could compare the information in all system outputs. The user was given the very simple instructions: "Rank these passages from best to worst" with no specification of the method they should use for the rankings. Users were also told they could indicate ties between 2 or more items for a particular question if they felt there was no qualitative difference. In the first experiment, 4 Arabic MT systems were ranked using a test set that consisted of 50 questions: 25 with the capitalization and punctuation produced by the system and 25 in which orthographic information was stripped from all passages. The passages were randomly ordered for each question to minimize the bias toward or against any particular system based on placement. In a second experiment, 3 Mandarin MT systems were compared; because one system only produced output with no capitalization or punctuation, the test set consisted of 25 passages in which orthographic information was stripped from all passages. The real-time audio processing for our experiments was performed by the BBN AudioIndexer system, described in detail in [2], which produces a word/character error rate of roughly 20-30% for English, Arabic, and Mandarin news sources. Systems MT1 and MT4 in the Arabic experiment were commercial rule-based systems with translation memory; systems MT2 and MT3 were commercial statistical MT systems.

Figure 1 shows an example of one of the Arabic test passages from our evaluation. All four passages are clearly errorful MT output, and none can be read to determine simply the meaning of the source; however, there are several factors that can be used to compare the quality of the four passages. For example, three passages contain the words "Arabic" (or "Arab") and "Tunisia," providing evidence that the passage pertains to Arabic cities and Tunisia. Output passage P2 contains neither of these words, which suggests either that P2 is missing important information or that the other three systems inserted the same spurious information. Similarly, there is clearly a list of countries at the end of each passage, but only output passage P3 contains the full list "Denmark, Greece, Angola, Canada, France," each of which is present in at least one other output. This indicates that P3 contains more complete information than any other. Outputs P1 and P4 both contain "distraction" words (pregnant, compeer) that adversely affect the perception of the passage quality. By focusing on contrastive factors such as these, it is possible for a user to decide on a relative ranking of the passages presented. While we recognize that the absolute information content of each passage cannot be determined without access to the original spoken language, one of our goals in this work was to determine which ranking factors users felt were most important in their determinations of relative quality.

P1: Five Arabic cities of he Tunisia which chose the group call inviting government with the country nor annexes all from the Denmark and the Greek and Angolan and pregnant Canada and France
P2: Five cities are chosen by Group calling the government with the host country and not include all of the Denmark and Greece , Canada , France
P3: Five towns in the Arab world is Tunisia that have chosen the advocates of the Government with the host country does not include both Denmark, Greece, Angola, Canada, France
P4: Five Arabic cities Tunisia that chose the group call a government with the country the host adjoin all from Denmark Greece and as the appellation carrier compeer France

Figure 1: A Sample Evaluation Item: "Rank These Passages from Best to Worst"

Another factor in perceived quality of language data, such as ASR and MT output, is the presence and quality of orthographic features such as capitalization and punctuation. Accurately-placed orthographic features, such as the capitalization of location names in Figure 1, can improve the readability of a passage.

P1: feared that al-qaeda threats that intelligence information became available in new york and washington which wnywjrsy
P2: afraid of threats which the organization which rule out the available information concerning its intelligence in washington york and
P3: frighten are evolved its al-qaida organization threats that and abounded the information the intelligence in her respect in washington new york new jersey
P4: fears that formation al-qaeda have fun threats of organization and which abounded intelligence information in his regard in washington and new york wnywjrsy

Figure 2: A Sample Evaluation Item with Orthographic Features Removed

Poorly-placed, inconsistent, or missing features, such as the lack of punctuation in passages P1 and P4 or the capitalized "Group" in P2 in Figure 1, similarly can hurt readability. In our experiments we sought to determine the contribution of orthographic features to the relative rankings of the MT systems. Figure 2 shows an example of a test passage with orthographic information removed.

5. RESULTS

The MT ranking experiment was completed by 8 different system users with a range of backgrounds. Education level of users ranged from no college degree to PhD. Linguistic background ranged from monolingual English to multilingual (but no Arabic or Mandarin), including two non-native English speakers. Users reported that the time required to complete 50 test questions ranged from 2 to 4 hours.

5.1. Rank Distribution

The relative ranking of the Arabic systems was nearly identical for all users, with one clear winner and two clear losers. Figure 3 shows histograms of the ranks assigned to passages from each of the 4 MT systems by the users. System MT3 was ranked 1^{st} in 213 of the 391 (54.5%) passages and either 1^{st} or 2^{nd} in an overwhelming 313 of 391 (80.1%). MT2 received half as many 1^{st} place votes as MT3 but was ranked either 1^{st} or 2^{nd} in 65% of the passages. In contrast the ranks of passages from systems MT1 and MT4 had nearly identical histograms, with very few 1^{st} place votes and 60-65% 3^{rd} or 4^{th} place ranking.

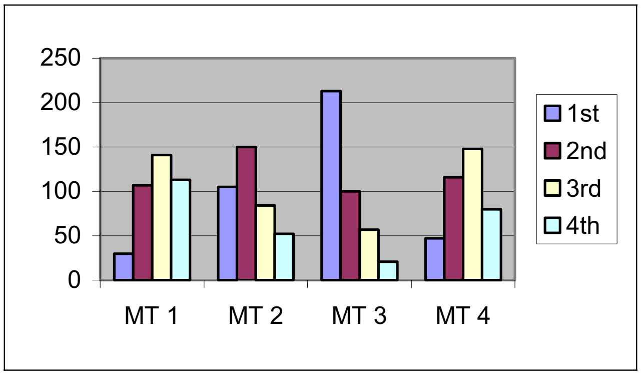

Figure 3: Histogram of Ranks Assigned to Arabic-to-English MT System Output Passages

The average rank assigned by users to MT3 passages was 1.82 (of 4), while the average ranks for MT1, MT2, and MT4 were 3.01, 2.35, and 2.82, respectively. The clear consensus from the 8 system users was that system MT3 was the best, MT2 was the second best, and MT1 and MT4 were almost interchangeably third and fourth.

We repeated the ranking experiment with a comparison of the 3 commercial Mandarin systems: two statistical MT systems and one rule-based with translation memory. The users showed a preference for two of the systems (the rule-based and one of the statistical), ranking them almost interchangeably 1st and 2nd, while ranking the other consistently 3rd. The average ranks for the three Mandarin systems were 1.66, 1.74, and 2.60 (of 3).

5.2. User Ranking Criteria

At the end of the experiment, users were asked to summarize the criteria they had used in their rankings. Users reported very different criteria, which makes the consistency in the relative rankings more significant. In order of reported importance, the criteria were:

Readability: While users gave better ranks to passages that they were able to read easily, the majority of the users reported that they gave up almost immediately trying to read the passages as a whole. Users instead developed strategies for identifying anchor words or phrases that could be used for direct comparison of the passages.

Name omission: Name phrases were important anchors for users in their comparisons, and users gave worse ranks to passages that did not contain names that were in the majority of other passages.

Numbers: Numbers were also important anchors for direct comparison. Users assigned worse ranks to passages containing incorrect or awkward number phrases, such as “five churches four” (vs. “fifty four churches”).

Distraction words: Users gave worse ranks to passages that contained words that distracted from the information content, such as “ossicle” and “rhymester.”

Passage length: Users assigned worse ranks to passages that were noticeably shorter or longer than all others.

Ties: Users very frequently assigned the same rank to 2 or more completely unintelligible passages within a set.

5.3. Orthographic Information

The average rank of each Arabic system differed slightly when the orthographic information was considered, showing that user perception of the MT quality was dependent not just on the words in the passage but also on the capitalization and punctuation. The average rank improved for two systems, indicating that the orthographic information made the system output more readable to the users. However, the average rank degraded (increased) for one system, indicating that this information made its output less readable compared to the other systems. The average rank of System MT3 improved from 1.9 (with no orthographic information) to 1.7 and MT4 from 2.9 to 2.7, while the average rank of MT2 degraded from 2.2 to 2.6.

6. SUMMARY

Our preliminary results suggest that user-centered experiments provide a useful comparison of MT systems from a real-world perspective. Users with a wide range of educational and linguistic background produced very similar relative rankings of 4 Arabic and 3 Mandarin MT systems. The user ranking method has the distinct advantage of eliminating the need for reference transcripts, which can be particularly difficult to prepare from spoken language input. Users participating in the ranking experiment indicated that MT of ASR output is still far from being easily readable, yet they were able to produce a very clear consensus on the relative quality of the MT systems being evaluated.

While we are encouraged by the results of the initial ranking experiments, there are many possible directions for further experimentation. The user ranking method of evaluation provides a good comparison of the relative usefulness of MT system outputs for a real-world task, but it does not address the adequacy dimension of MT evaluation. For example, similar insertion errors by multiple MT engines can provide a misleading view of the actual content of the source passage. It would be interesting to include a reference translation in the set of candidates presented to users for ranking, to see if the additional information in the reference changes the system rankings. Preparing this reference would require the resolution of the questions raised in Section 2 regarding ASR errors in the input transcript. One option would be to extend the ranking paradigm to bilingual evaluators, who would be presented with the set of target language MT outputs, as well as the source language ASR passage that was input to the MT system. Another possibility would be to provide bilingual users with access to the synchronized audio playback, so that they could hear the original spoken language for comparison with the ASR transcript and the MT system outputs.

7. REFERENCES

[1] M. Federico, “Evaluation Frameworks for Speech Translation Technologies,” In *Proc. of Eurospeech 2003*, pp. 377-380, 2003.

[2] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, “Speech and Language Technologies for Audio Indexing and retrieval,” In *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338-1353, 2000.

[3] D. Palmer, P. Bray, M. Reichman, K. Rhodes, N. White, A. Merlino, F. Kubala. “Multilingual Video and Audio News Alerting,” In *Proceedings of HLT/NAACL 2004 Demonstrations,* 2004.

[4] M. Paul, H. Nakaiwa, and M. Federico, “Towards Innovative Evaluation Methodologies for Speech Translation,” *Working Notes of NTCIR-4*, Tokyo, June 2004.

[5] K. Papineni, S. Roukos, T. Ward & w.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” In *Proceedings of the 40th ACL*, pp 311-318, Philadelphia, 2002.

[6] J. White & T. O'Connell, “The ARPA MT evaluation methodologies: evolution, lessons, and future approaches,” *Proceedings of the 1994 Conference, Association for Machine Translation in the America,* 1994.