Machine Translation Project Proposal

By: Ksenia Sokolova, Amelia Wolford
John Nunes, and Hyerim Ko

I.Summary of Pilot Project Outcomes

In total, 10 successful rounds of training were completed (see Appendix 1 for the full Pilot PEMT Project Workflow) with an initial selection of four bidirectional PDFs totaling approximately 10,000 segments as training data. Tuning and testing data sets initially consisted of two and three bidirectional PDFs respectively. The first round of training produced a BLEU score of 23.22, which over the course of the two-week pilot project increased by 5.78 points (or by 25%) to give a BLEU score 28.02 by the final round. The initial strategy, used in the first six successful rounds of training, was to vary the filetype used for training, tuning, and testing documents to determine the best possible choice: .DOC, .TXT, and .TMX files were used, with varying results. In the remaining rounds of training, a dictionary was added and later expanded, additional bi- and monolingual documents were added to the training data set, and various filetype combinations were tested.

TRAINING #	BLEU SCORE	PRODUCTIVITY points	PRODUCTIVITY %
1	23.22		
2	failed		
3	17.39	**-5.83**	**-25**
4	30.31 mistake	12.92	74
5	21.22	**3.83**	**22**
6	29.62	**8.4**	**40**
7	22.53	**7.1**	**-24**
8	failed		
9	failed		
10	failed		
11	23.10	**0.6**	**3**
12	27.30	**4.2**	**18**

13	30.27 mistake	2.97	11
14	failed		
15	28.02	**0.7**	**3**
16	29	**0.98**	**3.5**
ME training **productivity gain**			**25**

II. Proposed Objectives

In our pilot project proposal we predicted PEMT to be 25% more efficient than HT in terms of time and cost. The following tables demonstrate that these projected target criteria have been met and exceeded. The data collected during ME training allow us to expect possible increase in time/cost savings for continued training project of up to 30%.

Efficiency (1000 word sample)

	PEHT	PEMT
words/hr	**382**	**816**
time for 1000 words	2.6	1.2
time for review at 1000 words/hr	1	1
total time	3.6	2.2
time saving with PEMT	**1.4 = 39%**	

Cost (1000 word sample)

	PEHT	PEMT
translation rate	$0.12/word	$0.09/word
subtotal	$120	$90
review rate	$0.06/word	$0.06
subtotal	$60	$60
total	$180	$150
cost saving with PEMT	**37.5%**	

These data show that our system has already proven its cost- and time-saving ability.

Quality

After stating in our pilot project proposal that the TAUS machine translation quality guidelines would be used to evaluate PEMT quality, we determined that this method was not sufficiently objective and otherwise unsuitable for our use. Instead we have assessed PEMT quality based on the SDL LISA QA model.

In the LISA QA metric, errors are categorized as Minor, Major or Critical. Scores for all segments in the task are added together to give a total score for the task. If the overall quality of the task falls below a predefined threshold, the task will fail the LISA check.

	Minor	Major	Critical
Doc Language			
Mistranslation	1	5	10
Accuracy	1	5	10
Terminology	1	5	10
Language	1	5	10
Style	1	5	10
Country	1	5	10
Consistency	1	5	10
Doc Formatting			
Layout	1	5	10
Typography	1	5	10
Graphics	1	5	10
Call Outs and Captions	1	5	10
TOC	1	5	10
Index	1	5	10
Software Formatting			
Graphics	1	5	10
Alignment	1	5	10
Sizing	1	5	10
Truncation/Overlap	1	5	10
Character Formatting	1	5	10

Software Functionality Testing			
Localizable Text	1	5	10
Dialog Functionality	1	5	10
Menu Functionality	1	5	10
Hotkeys/Accelerators	1	5	10
Jumps/Links	1	5	10

Scorecard from LISA QA model by SDL: http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/LISA_QA_Model.htm

In medical translation quality is paramount. Precise medical translation is as crucial as the amount of anesthetic given to the patient before a surgery. Where safety depends on clear communication, comprehension of directions as well as medical device operation, accurate medical translations are key to ensuring that patients' safety is not put at risk.

In order to assess the quality of PEHT vs PEMT, both HT and MT were rated using the LISA QA model for error types with scores representative for a two hour translation/PE session. To pass the QA, translations must have no critical errors in any category and no major errors for mistranslation, accuracy and terminology. The number of errors in other categories must not exceed the score of 5. The overall score of >5 will fail the QA. LISA QA scores for PEHT VS PEMT content acquired during the project can be found in the Appendices (see Appendix 2).

QA provided in the Appendices (see Appendix 2) show that the two post-editing samples achieved a score of 2 (for PEHT) and 5 (for PEMT) respectively. In both cases this does not exceeds the maximum passing score of 5. Due to the expected increase in productivity/time/cost savings further ME training should be approached with quality as the primary goal in mind.

III. Recommendations for Continued Training

Based on the summary of the pilot project results, we can make several recommendations for further training the SMT engine:

- Continue training using texts related to medical devices. Currently the majority of documents relate to ultrasound devices and other imaging software and equipment,

and acceptable results have been achieved with this limited range of documents. However if the client would like to use the MT system for more subtopics within the broader domain of healthcare, then a wider variety of documents should be added to the training data set, keeping in mind that in order to maintain quality given a greater variety of data the number of documents will necessarily have to be increased.

- Continue training using only PDF and TMX. The group has discovered that Microsoft Hub prefers to work with our original PDFs. While we expected other formats (.doc, .txt) to yield better results, they did not. For this reason we recommend training with only PDF and TMX files, especially if time is limited. While during the pilot the use of TMX files at times resulted in a dramatic increase in the BLEU score and at other times led to a failed system, we expect the score to actually improve in the full scale project once proper time is invested in aligning the files (i.e. having someone align the files by hand vs automatic alignment tools which only allowed for a limited increase in the BLEU score).

- Add more documents to the training data set. The total number of training segments should be at least 100,000 segments (the minimum for full SMT engines). After several rounds of success using 4 bilingual texts as the training dataset, the group decided to add approximately 10,000 more segments to the training data set. After doing so, the BLEU score jumped. Adding more segments to the training data helped improve the score more than our other methods such as converting files, adding a dictionary, and "cleaning up the document."

- Adjust the mix of training and tuning documents. During the pilot project, only a few possible combinations were tested due to time constraints. Once we discovered a combination that yielded acceptable results we stuck with it. For training a full SMT engine we recommend continuing to experiment with the mix of the training and tuning dataset.

IV. Timeline and Costs

Due to the number of current unknowns, it is difficult to establish a true estimate of what it would cost to fully train an MT system. The main idea is that this is a long-term investment, and at least one full time employee would be required to monitor training.

V. Anticipated Results

Based on the results of the pilot project, a fully developed SMT engine is expected to yield equivalent or better results in terms of training productivity, quality, efficiency, and cost. In our pilot project proposal we predicted PEMT to be 25% more efficient than HT in terms of time and cost and exceeded this goal, demonstrating that the cost and time savings of 37.5% and

39% respectively compared to HT are achievable. While these results are expected to remain constant or improve, the primary expectation is that quality of the MT will improve. This is not to say that the engine will not need human post editing - it always will especially due to the (medical) nature of these texts.