ARMY RESEARCH LABORATORY

Human in the Loop Machine Translation of Medical Terminology

by John J. Morgan

ARL-MR-0743 **April 2010**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0743 **April 2010**

Human in the Loop Machine Translation of Medical Terminology

John J. Morgan
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2010	Interim	1st quarter FY10

4. TITLE AND SUBTITLE
Human in the Loop Machine Translation of Medical Terminology

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
John J. Morgan

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-T
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-MR-0743

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This memorandum report outlines the Human in the Loop Translation process, which strives to increase translators' productivity using statistical machine translation (SMT). In this process, U.S. Army Research Laboratory (ARL) systems trained by open source SMT toolkits were placed in a feedback loop with human translators in Afghanistan to incrementally improve Dari translations of English medical training manuals and to create bilingual glossaries in the medical domain. Anecdotal evidence indicates that the quality of the machine translation drafts produced using this feedback loop process is high enough to assist human translators in translating training manuals in the medical domain. Automatic scores showed large improvements in translation quality as the loop progressed for a very narrow task, providing indirect evidence that SMT can benefit human translators even with small amounts of training data.

15. SUBJECT TERMS
Statistical machine translation, medical terminologies, human in the loop

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John J. Morgan
Unclassified	Unclassified	Unclassified	UU	20	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-1902

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Tables

Acknowledgments

The Human in the Loop project was conceived and directed by Multilingual Computing Branch (MLCB) member Dr. Stephen A. LaRocca. I would like to thank Dr. LaRocca for his guidance and I would also like to thank Mr. Hazrat Ghulam Jahed for his work translating and editing.

INTENTIONALLY LEFT BLANK.

1. Introduction

The U.S. Army Research Laboratory (ARL) is looking for ways to support the mission of transitioning control of the war in Afghanistan from the United States and the North Atlantic Treaty Organization (NATO) to the Afghan government. In his speech at West Point, NY, on December 1, 2009, President Obama said, "We must strengthen the capacity of Afghanistan's security forces and government so that they can take lead responsibility for Afghanistan's future." This transition of control depends on the success of training the Afghan military. Success in training relies on effective communication between forces that speak different languages, which, in turn, relies on high-quality translation services. Although translators can produce high-quality work on their own, state-of-the-art machine translation technologies could provide translators many benefits, among them the ability to coordinate their efforts to archive and reuse translations.

This report describes how ARL's Multilingual Computing Branch (MLCB) is supporting the training mission in Afghanistan through the use of language translation technology:

1. The MLCB is working on the translation of a key training text. The techniques used in this translation can be used as a template for future translations of similar texts.
2. Corpora of in-domain texts are being archived and prepared for future processing by the MLCB and other research organizations including Afghan organizations.
3. Terminologies in the Afghan languages, Dari and Pashto, are being extracted and standardized.
4. The natural language processing (NLP) tools being used in this project are open source and freely available for use by Afghans in the future.

This report discusses the translation of one specific English document (a medical training manual) into Dari and describes a method for using machine-translated text with the aim of helping translators in the future.

2. Background

Currently, the U.S. Army is training soldiers in the Afghan National Army (ANA), including soldiers who aspire to become medical doctors and nurses. The trainers are using the same manuals, originally written in English, which are used to train U.S. Soldiers. The ANA would like to have these manuals available in the trainees' native languages, which, in the case of many of them, is Dari or Pashto. In order for the training projects to be sustained by Afghans in the

future, these training manuals will need to be available in the languages of the Afghan people. Additionally, methods for translating new training documents should also be available.

In fall 2008, the MLCB received a request for help from a military medical doctor working with an Embedded Training Team (ETT) as part of the Combined Security Transition Command-Afghanistan (CSTC-A). The physician needed a Dari translation of an English-language medical training manual, *Fundamental Critical Care Support (FCCS)*, published by the Society for Critical Care Medicine (SCCM) and used as a textbook for training doctors and nurses for work in Intensive Care Units (ICUs). Knowing that the MLCB worked on language technology, the military doctor wanted to know if machine translation could help with creating such a translation.

Of interest was whether a machine translation of a medical document from English into Dari could assist human translators in translating the same document. Human translators on the ETT can translate ~3 pages per day. Machine translation speed depends on computer hardware, but it is reasonable to expect that machine translation systems on a current desktop computer can translate 1 sentence per second. Thus, machine translation could produce a human translator's daily output in 1 min. However, even the best machine translations are usually not acceptable to human readers. One can only expect a rough draft translation from a computer-generated translation. Current state-of-the-art machine translation systems rely on large parallel corpora (databases of human-translated texts) and NLP tools. Work is just beginning on collecting these corpora and developing the NLP tools for the languages of Afghanistan.

Given the physician's access to a team of highly skilled translators who were familiar with medical terminology in both English and Dari, the MLCB proposed a partnership with the ETT's team of translators to develop a new translation process. In this process, termed "Human in the Loop Translation," the MLCB provides rough draft machine translations of the FCCS text into Dari; the human translators edit the rough draft and return the edited copy to the MLCB; and then the MLCB updates the machine translation corpus based on the translators' changes. Partnering in this way yields potential benefits to both the ETT and our lab. The military medical doctor increases the productivity of his team of translators, while the MLCB collects an English/Dari bilingual parallel corpus in the medical training manual domain, which could then be used to translate future training manuals and also serve as a source of data for scientific research.

The Human in the Loop process also addresses another issue that doctors working on ETTs brought up: they felt that they were starting all over again every time a new team arrived. An archive of translated documents and bilingual terminologies would remedy this problem, further enabling the transition of future medical care to Afghan responsibility.

3. Statistical Machine Translation (SMT) Systems

NLP tools and other language resources are very scarce for Dari. Specifically, English/Dari bilingual text corpora were almost nonexistent when the command surgeon contacted the MLCB. We proposed the Human in the Loop process in order to build a machine translation system that would gradually improve as more domain-specific bilingual text became available for training the system. For that task, we considered using open source, statistical machine translation (SMT) systems. These systems have the following advantages:

1. They provide the means to create complete translation systems. These packages include a decoder, which actually performs the translation by taking an English sentence as input and searching for the best Dari translation to output.
2. In most cases, they provide scripts for training the components the decoder needs to perform translations.
3. They only require a parallel corpus of text to produce an effective system. They do not require experts in any field to produce grammars, mathematical models, or any complicated software engineering.
4. The systems can be built and rebuilt rapidly once the training corpus is available. Training one system on our 50,000 segment parallel corpus takes two days on a single processor. Parts of the process of building these systems can be parallelized and run on a cluster of computers to speed up the process as the training corpus grows.

We explored two options for training SMT systems: Moses (*1*), which implements a phrase-based translation method; and the Johns Hopkins open source architecture (JosHUa) (*2*), a new third-generation SMT system. Broadly speaking, first-generation systems translate words, second-generation systems translate phrases, and third-generation systems translate structures. Specifically, JosHUa implements a system called Hiero (*3*) that automatically extracts grammar structures called synchronous context free grammars (SCFGs) and decodes using chart parsing.

The field of machine translation seems to be moving toward the syntax-based approach of JosHUa and our results show better performance from JosHUa than Moses.

4. Human in the Loop Translation Process

First, the systems were trained on a bilingual corpus that we had been aligning at the MLCB for over a year. The source of the corpus was a newspaper published in Kabul, Afghanistan, in the three languages English, Dari, and Pashto. We also had access to English to Dari translations of Army field manuals that had been done recently.

After resolving copyright issues with the SCCM, we received the entire FCCS book in portable document format (pdf). After extracting the text from the pdf file, we ran the first chapter through our systems trained on the newspaper and field manual corpora. The translations were very rough; many technical medical terms were not translated since they did not appear in our training corpus. A two-column spreadsheet was composed, the first column containing each line of the chapter in English and the second column containing the machine-rendered translation in Dari. The spreadsheet was e-mailed to the military physician in Afghanistan, and the translators wrote their version in a third column of the spreadsheet and sent it back to the MLCB. The first and third columns of the spreadsheet were extracted, and the systems were retrained with this new set of ~250 English/Dari segments. We continued this process on subsequent chapters.

As of this report, chapters 1 through 6 have been translated and chapter 7 has yet to be sent back from the translators in Afghanistan. At the end of this project, we envision publishing the entire FCCS book in an English/Dari bilingual printed book format, where facing pages contain mirrored translations: English on the left-hand side and Dari on the right-hand side. Although a printed book may seem low tech, this format is what trainers in the field have requested and what they feel will be most helpful in their work with trainees.

Trainers and other translation professionals have also expressed an interest in bilingual (English/Dari) glossaries in special domains. The remainder of this report will focus on extracting and translating medical terminology from English into Dari.

We have extended our previous work to a Human in the Loop procedure for creating a bilingual English/Dari glossary of medical terms similar to the method used by Deléger, Merkel, and Zweigenbaum (*5*). According to these researchers, medical terminology is a necessary resource for any kind of health care information task (e.g., coding, free text indexing, information retrieval). Extracting terms from text can be very tedious. If terms only consisted of single words, the task of terminology extraction would be easy: a human would be given a list of words and prompted to decide whether or not the word is a term. But terms are not limited to single words. Thus, a human has to read the text in context in order to determine the technical terms.

Deléger, Merkel, and Zweigenbaum worked with the heavily resourced English-French language pair, so their procedure was slightly different. Once they extracted their terms, they displayed a sentence from a parallel corpus in which it occurred in context.

4.1 Medical Language System (UMLS) Tools

In our procedure, we leverage work done by the Lexical Systems Group (LSG) at the National Institutes of Health (NIH) in creating the Unified Medical Language System (UMLS) (*12*). LSG has developed tools for extracting technical terms from text in the biomedical domain. We used these tools to extract English terms from the FCCS. Specifically, we took the following steps with the UMLS tools:

1. Text was parsed into phrases using the Java class *gov.nih.nlm.nls.nlp.parser.Parse*.
2. Next, noun phrases and prepositional phrases were extracted (with prepositions stripped).
3. Terms were extracted using the LSG tool lexical variant generator (lvg). We experimented with different options and flow components for lvg. The following is one example of a command line for lvg: **lvg -f:Ct:fa:g:p**. In this program, the Ct option retrieves the unique lexical names of the terms, the fa option filters out acronyms, the g option removes genitives, and the p option strips punctuation.
4. The list of terms was fed to the Moses or JosHUa decoders, which wrote out candidate Dari translations.
5. The candidate translations were edited by a native Dari speaker.
6. Finally, the edited translations were appended to the training corpus.

So far, the Moses and JosHUa SMT systems have been trained with ~65,000 segments from parallel texts in the newspaper, military training manual, and medical domains.

4.2 Training Moses and JosHUa

The training steps for the Moses and JosHUa systems included the following:

1. Sentence level alignment of the parallel training corpus was performed using Robert Moore's Bilingual Sentence Aligner (*6*).
2. Word level alignments were extracted using the Posterior Constraints Alignment Toolkit (PostCAT) (*7*) and the Berkeley Aligner (*13*).
3. The Java class *joshua.corpus.suffix_array.Compile* was used to put the word alignments into suffix array data structures, based on the syntax-based translation method developed by Adam Lopez (*11*). Note: Moses does not yet have the capability to implement this step.
4. In the JosHUa system, grammar rules were extracted from the suffix arrays and a development set of segments with the Java class *joshua.prefix_tree.ExtractRules*. In Moses, a table of phrase translations was extracted from the word alignments obtained in step 2.

5. The grammar rules in the JosHUa system and the phrase pairs in the Moses system were scored using relative frequency statistics.
6. Both systems used the Stanford Research Institute Language Modeling (SRILM) toolkit to make a 4-gram statistical language model for Dari.
7. A reordering model was trained for Moses. Since reordering is modeled by the grammar rules, JosHUa did not train a reordering model.
8. Both systems trained weights for the log-linear model with a minimum error rate training (mert). JosHUa used Omar Zaidan's zmert via the Java class *joshua.zmert.Zmert* (*9*) and Moses used the Franz Och original version of mert (*10*).

The log linear model was used to combine the feature functions including those defined by the language model and phrase table (in the case of Moses) or grammar rules (in the case of JosHUa).

5. Results and Discussion

The output from the two systems was reviewed by a MLCB native Dari speaker. He found the output from the Moses system to be a useful aid to translators as a rough draft, saying that he would use them himself if he were asked to translate the FCCS. He also noted that the quality of the translations was improving with every cycle in the loop.

While this subjective evaluation was encouraging, a more objective look at the effect of our process on the quality of machine-translated text was needed.

The FCCS text includes a test that is given to trainees after instruction to assess their learning. This test, which includes terminology from the entire FCCS text, was translated into Dari. We then used the set of 311 segments in the test to evaluate the performance of our systems (table 1). We obtained Bilingual Evaluation Understudy (BLEU) (*12*) scores using version 12 of the National Institute of Standards and Technology's (NIST) MTeval utility (*13*) (ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl) on the output from the Moses and JosHUa systems. BLEU is a standard, albeit controversial, metric for evaluating the performance of machine translation systems. We used the BLEU scores here to confirm the trend we noticed anecdotally. The scores were conveniently obtained via MTeval and we do not make claims concerning accuracy of BLEU as a metric.

Table 1. BLEU scores for systems trained incrementally on chapters of the FCCS text.

Training	Moses BLEU	JosHUa BLEU
baseline	0.148	0.1687
baseline + chapter 1	0.1923	0.2127
baseline + chapter 1 + mert	0.2113	0.2345
baseline + chapters 1, 2	0.256	0.2678
baseline + chapters 1,2 + mert	0.2545	0.2929
baseline + chapters 1,2, 3	0.257	0.2799
baseline + chapters 1,2,3 + mert	0.246	0.3115
baseline + chapters 1,2,3, 4	0.2868	0.3029
baseline + chapters 1-4 + mert	0.2984	0.3366
baseline + chapters 1-5	0.2991	0.3209
baseline + chapters 1-5 + mert	0.2766	0.3528
baseline + medical domain + chapters 1-5	0.3751	0.3830
baseline + medical domain + chapters 1-5 + mert	0.327	0.4105

The baseline system was trained on text from the newspaper and field manual domains. Our medical domain data were excluded from the baseline system's training corpus.

For the JosHUa system, appending a new chapter to the training corpus clearly resulted in improved BLEU scores on the test set, although in three cases running mert yielded better BLEU scores than appending the next chapter's text to the training corpus. BLEU scores also rose with each new chapter appended to the Moses training corpus, but mert training lowered BLEU scores in four cases.

The FCCS contains 15 chapters. Table 1 shows that the Human in the Loop procedure increases the performance of an SMT system based on the BLEU scores in the upper teens to the upper 30s, even before translating half the chapters in the book.

6. Conclusions

Trainers who need translations of specific texts and in a very specific domain, like critical care medicine, can benefit from automatically generated machine translations by participating in a Human in the Loop process. The systems reported on in this report were trained on small amounts of data, but by periodically updating their training corpora, the systems eventually produced useful translations. In translating a book, for example, we believe that using this procedure would yield acceptable results by the time half the chapters in the book were translated.

7. References

1. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, June 2007, pages 177–180, Association for Computational Linguistics. http://aclweb.org/anthology-new/P/P07/P07-2045.pdf.

2. Li , Z.; Callison-Burch, C.; Dyer, C.; Ganitkevitch, J.; Khudanpur, S.; Schwartz, L.; Thornton, W.N.G.; Weese, J.; Zaidan, O. F. Joshua: An Open Source Toolkit for Parsing-based Machine Translation. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 30–31, 2009, http://www.aclweb.org/anthology/W/W09/W09-0x24.

3. Chiang, D.; Lopez, A.; Madnani, N.; Monz, C.; Resnik, P.; Subotin, M. The Hiero Machine Translation System: Extensions, Evaluation, and Analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, October 6–8, 2005, 779–786, http://dx.doi.org/10.3115/1220575.1220673, Association for Computational Linguistics: Morristown, NJ.

4. Deléger, L.; Merkel, M.; Zweigenbaum, P. Translating Medical Terminologies Through Word Alignment in Parallel Text Corpora. *J. of Biomedical Informatics* **August 2009**, *42* (4), 692–701.

5. Moore, R. C. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine *Translation: From Research to Real Users*, Richardson, S., ed.; Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, CA, 135–244, 2002, http://link.springer.de/link/service/series/0558/bibs/2499/24990135.htm

6. Graça, J.; Ganchev, K.; Taskar, B. Post-cat - Posterior Constrained Alignment Toolkit. *The Third Machine Translation Marathon* **2009**, *91*, 27–36.

7. Stolcke, A. SRILM an Extensible Language Modeling Toolkit. *Intl. Conf. on Spoken Language Processing*, 2002, 901–904.

8. Zaidan, O. F. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics* **2009**, *91*, 79–88.

9. Och, F. Jf. Minimum Error Rate Training for Statistical Machine Translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, Sapporo, Japan, July 7–12, 2003, 160–167, http://www.aclweb.org/anthology/P03-1021.

10. Lopez, A. D.; Resnik, P. S. *Machine Translation by Pattern Matching*; University of Maryland at College Park, College Park, MD, 2008, ISBN: 978-0-549-57255-8.

11. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research* **January 2004**, *32*, Database issue D267–D270.

12. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. *BLEU: A Method for Automatic Evaluation of Machine Translation*; Technical Report RC22176 (W0109-022); IBM Research, 2001, 311–318.

13. Doddington, G. Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrence Statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 2002.

14. Liang, P.; Taskar, B.; Klein, D. Alignment by Agreement. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, NY, June 4–9, 2006, 104–111 [doi>10.3115/1220835.1220849].

List of Symbols, Abbreviations and Acronyms

ANA	Afghan National Army
ARL	U.S. Army Research Laboratory
BLEU	Bilingual Evaluation Understudy
CSTC-A	Combined Security Transition Command – Afghanistan
ETT	Embedded Training Team
FCCS	Fundamental Critical Care Support
ICU	Intensive Care Unit
JosHUa	Johns Hopkins open source architecture
LSG	Lexical Systems Group
lvg	lexical variant generator
mert	minimum error rate training
MLCB	Multilingual Computing Branch
NATO	North Atlantic Treaty Organization
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NLP	natural language processing
pdf	portable document format
PostCAT	Posterior Constraints Alignment Toolkit
SCCM	Society for Critical Care Medicine
SCFG	synchronous context free grammar
SMT	statistical machine translation
SRILM	Stanford Research Institute Language Modeling
UMLS	Unified Medical Language System

NO. OF
COPIES ORGANIZATION

1 PDF ADMNSTR
DEFNS TECHL INFO CTR
ATTN DTIC OCP
8725 JOHN J KINGMAN RD STE
0944
FT BELVOIR VA 22060-6218

17 HCS US ARMY RSRCH LAB
1 PDF ATTN RDRL CII
B BROOME
ATTN RDRL CII T
J J MORGAN (12 HCs, 1 PDF)
ATTN RDRL CII T
M HOLLAND
ATTN RDRL CIM P
TECHL PUB
ATTN RDRL CIM L
TECHL LIB
ATTN IMNE ALC HRR
MAIL & RECORDS MGMT
ADELPHI MD 20783-1197

1 HC US ARMY RSRCH LAB
ATTN RDRL CIM G T LANDFRIED
BLDG 4600
ABERDEEN PROVING GROUND MD
21005-5066

TOTAL: 20 (18 HCs, 2 PDFs)

INTENTIONALLY LEFT BLANK.