

ARMY RESEARCH LABORATORY



Project-specific Machine Translation

by John J. Morgan

ARL-TR-5854

December 2011

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-TR-5854

December 2011

Project-specific Machine Translation

John J. Morgan

Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) December 2011		2. REPORT TYPE Final		3. DATES COVERED (From - To) FY11	
4. TITLE AND SUBTITLE Project-specific Machine Translation				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-TR-5854	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Virtual Translation Center Suite NVTC-200 935 Pennsylvania Ave. N.W. Washington, DC 20535				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 report details a project-specific method for translating a book-length text. The method used a human-in-the-loop strategy to iteratively and incrementally produce a progressively more accurate statistical machine translation system. The system generated rough-draft translations of chunks of text that were post-edited by an expert translator. The method was used to produce a high-quality translation of the Army Field Manual (FM) 7-8 "Infantry Rifle Platoon and Squad" from English into the Afghan language Pashto. The results show that through a process of over-fitting, this method leads to systems that are optimized for aiding a single translator.					
15. SUBJECT TERMS Machine translation, low resource language					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 38	19a. NAME OF RESPONSIBLE PERSON John J. Morgan
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (301) 394-1902

Contents

List of Tables	v
1. Introduction	1
1.1 Problem Description.....	1
1.2 The Mission.....	1
1.3 The Team.....	1
1.4 Definition of a Project.....	1
1.5 Steps in the Project.....	2
1.6 Over-fitting.....	3
1.7 Previous Work in Incremental System Rebuilding.....	3
1.8 Evaluation.....	4
2. Methods	4
2.1 Corpora.....	4
2.1.1 Baseline Corpus.....	5
2.1.2 FM 7-8 Corpus.....	5
2.1.3 Ranger Handbook Corpus.....	5
2.2 Comparison of Chapter Titles.....	6
2.3 Pre-editing.....	6
2.4 Translation Models.....	7
2.5 Training Outline.....	7
2.5.1 Language Models.....	7
2.5.2 Word Alignment.....	8
2.5.3 Translation Tables.....	8
2.5.4 Tuning.....	8
2.6 Evaluation.....	8
3. Results	8
3.1 Baseline on FM 7-8.....	8
3.2 Baseline on Ranger Handbook.....	9
3.3 Stages.....	9
3.3.1 FM 7-8 on FM 7-8 by Stage.....	10

3.3.2	Ranger Handbook on Ranger Handbook by Stage.....	10
3.3.3	Ranger Handbook on FM 7-8 Results by Stage	11
4.	Conclusion	11
4.1	Conclusions	11
4.2	Future Work	11
5.	References	13
	Appendix A. Corpus Statistics	15
	Appendix B. System Scores	19
	List of Symbols, Abbreviations, and Acronyms	29
	Distribution List	30

List of Tables

Table 1. Numbers of bi-segments used in training the baseline system.	5
Table 2. Comparison of chapter and section titles between FM 7-8 and the Ranger Handbook.	6
Table 3. BLEU scores for the baseline system on each chunk of text from FM 7-8.	9
Table 4. BLEU scores for the baseline system on each chunk of text from the Ranger Handbook.	9
Table 5. FM 7-8 BLEU scores for test of systems at each stage. The first column shows the text used in training and tuning. The second column indicates the test text. The third column shows the BLEU score for the test.	10
Table 6. Ranger Handbook BLEU scores for test of systems at each stage. The first column shows the text used in training and tuning. The second column indicates the text. The third column shows the BLEU score for the test.	10
Table 7. BLEU scores for systems trained and tuned on Ranger Handbook text and tested on FM 7-8 at each stage. The first column shows the text used in training and tuning. The second column indicates the text. The third column shows the BLEU score for the test.	11
Table A-1. Number of tokens in each corpus.	15
Table A-2. Number of bi-segments from chunks of FM 7-8.	15
Table A-3. Cumulative number of segments from FM 7-8.	16
Table A-4. Total number of types in each training stage. Notice how slowly these numbers increase.	16
Table A-5. Number of tokens in each chunk of FM 7-8.	17
Table A-6. Cumulative number of tokens from FM 7-8 by stage.	17
Table A-7. Numbers of bi-segments in each chapter of the Ranger Handbook.	18
Table B-1. All FM 7-8 BLEU scores for systems by stage and FM 7-8 chunk.	20
Table B-2. All results of FM 7-8 system tested on Ranger Handbook.	22
Table B-3. BLEU scores for Ranger Handbook trained and tuned systems tested on Ranger Handbook.	24
Table B-4. BLEU scores for tests of systems trained and tuned on Ranger Handbook data and evaluated on FM 7-8.	26
Table B-5. BLEU scores for two systems at stage IX: one with tuning and one without tuning.	27
Table B-6. BLEU scores for two stage-X systems: one with tuning and one without tuning.	28

INTENTIONALLY LEFT BLANK.

1. Introduction

This section outlines the what, why, who, and how regarding a project-specific method for translating a book-length text. The method used a human-in-the-loop strategy to iteratively and incrementally produce a progressively more accurate statistical machine translation (SMT) system. The system generated rough-draft translations of chunks of text that were post-edited by an expert translator. The method was used to produce a high-quality translation of the Army Field Manual (FM) 7-8 “Infantry Rifle Platoon and Squad” from English into the Afghan language Pashto.

1.1 Problem Description

Typical SMT systems are built with large amounts of text data in the form of bi-segments—pairs of sentences in the source language and their corresponding translations in the target language. However, for languages like Pashto, few bi-segment data resources are available. This lack of resources is referred to as “data paucity.”

A project-specific method is described here to translate a book-length text into a language with severe data paucity and few natural language processing resources (i.e., Pashto).

1.2 The Mission

The method was implemented in the context of a project to assist linguists in translating Army FM 7-8 (*13*) from English into Pashto. This was part of a mission taken on by the Multilingual Computing Branch (MLCB) at the U.S. Army Research Laboratory (ARL) to provide security assistance to the Afghan National Army (ANA).

1.3 The Team

The team consisted of four people. Two of the team members were bilingual English-Pashto speakers. The mission was lead by an English-speaking military subject matter expert (SME). An English-speaking language technologist developed the SMT systems. One of the bilingual team members served as the main translator and military SME in the Pashto language. The other bilingual member served in a coordinating role for the team, working separately but closely with both the SMT developer and the translator.

1.4 Definition of a Project

In its general sense, the concept of “project” usually entails a temporary endeavor involving the following:

1. a goal, such as a customer with a need or an institution with a requirement;

2. a timeframe or plan, such as a sequence of tasks to be performed cooperatively within a given period established to achieve the goal; and
3. dedicated resources, such as skilled personnel, equipment, funding support, etc.

The National Virtual Translation Center (NVTC) presented a goal, requirements, a timeframe, and funding support, and ARL’s MLCB responded with a plan and dedicated resources. Thus, the term “project” was chosen to describe the method, which is outlined as follows:

1. The specific goal was to produce a high-quality translation of a lengthy text* despite severe data paucity.
2. The plan was a stepwise progression of coordinated tasks, including the following:
 - training and tuning SMT systems,
 - generating draft translations, and
 - editing drafts.
3. The timeframe was affected by the availability of resources and the urgency of the requirements. Resources included personnel with the necessary skills to support the performance of tasks:
 - a translator,
 - an SME for both the source and target languages,
 - a bilingual editor, and
 - a language technologist.

Algorithm 1 provides a brief sketch of the method used in this project:

Algorithm 1 A brief sketch of the method.

```

for  $i = 1 \rightarrow \approx 13$  do
  SMT  $\leftarrow$  draft
  human edit
  rebuild system
end for

```

1.5 Steps in the Project

The steps in the project can be summarized as follows:

1. The SMT system produces a rough draft translation of a chunk of text.

*A book-length text containing at least 3000 segments qualifies as lengthy.

2. The draft is post-edited by the human SME and the translator.
3. The language technologist rebuilds the entire SMT system with the new data.

The recently edited chunk of text gets used in three main ways in rebuilding the new system:

1. To train the translation model (TM)
2. To train a separate language model (LM) for the project's text.
3. As data for weight tuning with Minimum Error Rate Training (MERT) (8)

After each chunk of text is post-edited by the SME, each of the above steps are repeated.

On a machine with two quad core processors and 32 GB of memory, close to 20 h are required to complete each rebuilding stage.

During this project, the language technologist entirely rebuilt the SMT system with chunks of approximately 400 segments from the text about every two weeks. The baseline system was trained on around 33k bi-segments. Each stage of the rebuilding process should be thought of as a new system designed to translate the next chunk of text. The final system should be thought of as a system designed to translate the specific text, in this case FM 7-8. The question remains, does a book deserve its own translation engine?

1.6 Over-fitting

Basic researchers and commercial developers who seek to produce systems that can serve as general purpose translation machines usually strive to enlarge the set of input text for which their system produces a good translation. That is not the goal here. In fact, this project-specific method produces a system that should only be considered useful for one translation. The method exploits the over-fitting tendency of the machine learning algorithms used to develop SMT systems to produce a good translation for only one translator and one chunk of input text. This is as much as can be expected from systems constructed in severe data paucity conditions.

The draft translations produced by the SMT system presents the translator with his previous lexical and grammatical choices. If the previous choices are correct, the translator is more likely to choose to use them again by leaving them intact instead of rewriting the translation and expressing himself in a slightly different way. This is a mechanism for over-fitting.

1.7 Previous Work in Incremental System Rebuilding

An algorithm for re-training an SMT system with post-edited sentences was developed (2) by a Danish team in an effort to make an interactive aid for human translators. The re-training is almost fast enough to achieve the goal of interactivity. The translation model updating with the most recently post-edited sentence occurs in seconds.

Another project (6) developed a specialized Expectation Maximization (EM) algorithm for retraining a system for translating streams of text from the Web.

Although these methods deal with problems similar to the one dealt with in this report, they assume the existence of larger amounts of training data than those available for Pashto.

1.8 Evaluation

The goal of a “project” is to aid a human translator in translating one chunk of text, but does it help? If so, then by how much? Further, is it better than other methods? Definitive answers to these questions are not given here. However, evidence is provided to show that the sequence of systems models the translator’s characteristics.

In addition to serving as training and tuning data, the recently post-edited chunk of text is used to gage the performance of the system built in the previous iteration. To illustrate this performance measurement, the system rebuilding stages are labeled I, II, III, etc., through XII and the recently post-edited chunks of text data are labeled as a, b, c, etc., through l. System I is rebuilt with chunk a, II with b, III with c, and so on, through XII with l. Then, system I is tested on chunk b, II on c, III on chunk d, etc., through XI on l. The baseline system is tested on chunk a. The last stage does not have a test corpus; the tests are not objective. The test data are edited by the same person who edited the bi-segments that were used to train the systems that are being tested. Thus, the results indicate to what extent the translations are meeting the translator’s expectations.

This report documents the test results of one system as the building process moved through 12 stages. A pattern of improving test scores was observed as more in domain data was added to the training and tuning corpora. However, questions remain. At what point in this process will the SMT system become a useful aid to a human translator? Anecdotal evidence indicates that the translator rejects the first SMT-generated draft translation. Is there a minimum threshold of bi-segments under which the method fails? This report considers the case where that threshold is around 33k.

Section 2 describes in more detail the text corpora used for training, tuning, and testing and the way the Moses Toolkit (3) was used to perform the experiment. Section 3 presents the BLEU scores for tests run using Moses hierarchical models.

2. Methods

The sets of text data used in building the SMT systems are described first.

2.1 Corpora

A baseline SMT system was trained on a corpus consisting of seven subcorpora. All the corpora were segment aligned. No text from the military domain was used in training the baseline system.

Only text from either FM 7-8 or the Ranger Handbook (14) was used to train systems built in later stages on top of the baseline system. The subcorpora are described in terms of their size as measured by the number of segments. Segments mostly correspond to sentences, but also include titles, section headings, etc.

Next, the corpus used to train the baseline is described.

2.1.1 Baseline Corpus

The numbers of segments left after conditioning are shown in table 1. Since text conditioning sometimes drops some bi-segments, the numbers here might be less than the total number of segments in the original texts. These are the numbers of segments that actually get used in training the baseline system. The corpora labeled LDC come from the Linguistic Data Consortium’s Reflex Pack for Less Commonly Taught Languages (LCTL) (7). The corpora labeled Sada-e-Azadi (SeA) (11) and Afghan Recovery Report (ARR) (4) come from publications on the Internet. Qamoosuna (dictionaries) (QAMO) is an online English-Pashto dictionary. The Legal corpus includes the Afghan constitution. Notice that these corpora are very small compared to a database like EUROPARL (5), which contains over a million segments.

Table 1. Numbers of bi-segments used in training the baseline system.

Corpus	Segments
ARR	2975
LDC Elicitation	2672
LDC News	10016
LDC Phrasebook	1123
Legal	619
QAMO	834
SeA	15643
total	33882

More training corpora statistics are listed in appendix A.

2.1.2 FM 7-8 Corpus

The entire FM corpus consists of 5478 segments. The FM 7-8 text was chunked at the section boundary closest to a cut that would produce 400 segments. The chunks were used to iteratively build the new SMT systems. Details of the chunks are described in appendix A-2.

2.1.3 Ranger Handbook Corpus

The Ranger Handbook is a text in the same genre—military field manuals and domain, small infantry unit tactics—as FM 7-8. Thus, it is suitable as a control group to test the effects of the project specific method. Table A-7 in appendix A-7 shows the number of segments by chapter of the Ranger Handbook.

2.2 Comparison of Chapter Titles

In table 2, the titles of the chapters of the Ranger Handbook are displayed side by side with the titles of the section and chapter titles of FM 7-8 where the chunk cuts were made. Notice that although the titles do not align, there is overlap in the topics covered by the two texts.

Table 2. Comparison of chapter and section titles between FM 7-8 and the Ranger Handbook.

chapter 1 Doctrine	chapter 1 PRINCIPLES OF LEADERSHIP
section 2a OPERATIONS	Chapter 2 OPERATIONS
section 2b actions at danger areas	Chapter 3 FIRE SUPPORT
section 2c DEFENSE	Chapter 4 MOVEMENT
section 2d other operations	Chapter 5 PATROLLING
section 2e ARMORED VEHICLE SUPPORT	Chapter 6 BATTLE DRILLS
section 2f NUCLEAR, BIOLOGICAL, AND CHEMICAL OPERATIONS	Chapter 7 COMMUNICATIONS
section 3a PATROLLING	Chapter 8 ARMY AVIATION
section 3b AMBUSH	Chapter 9 WATERBORNE OPERATIONS
chapter 4 BATTLE DRILLS	Chapter 10 MILITARY MOUNTAINEERING
section 5a INFANTRY PLATOON TACTICAL STANDING OPERATING PROCEDURE	Chapter 11 EVASION\SURVIVAL
section 5b OFFENSE	Chapter 12 FIRST AID
section 5c air defense artillery	Chapter 13 DEMOLITIONS
	Chapter 14 RANGER URBAN OPERATIONS

2.3 Pre-editing

The text conditioning that was performed on the English and Pashto sides are listed in this section.

All the English text was downcased. Tokenizations were performed on the following Unicode character property classes:

- **marks**
combining marks and punctuation marks
- **symbols**
including math and currency symbols
- **control characters**
The bidi marks were deleted.

The following general text conditioning was performed on all the text:

- Expansion of tab characters to white spaces.
- Conversion of hard spaces to white spaces.

- Squeezing of white spaces.
- Removal of segment initial and final white space.

The following normalizations were performed on the Pashto side:

- Conversion of Arabic digits to equivalent Eastern Arabic-Indic digits.
- Conversion of English punctuation marks to equivalents in the Arabic character set.
- Conversion of Arabic characters in the presentation form B Unicode block to their equivalents in the Arabic block.

Also, segments with more than 99 tokens were dropped and segment pairs were made unique in the following sense: Segments from the English and Pashto sides were pasted into one bi-segment file. The UNIX tool `sort` was run with the `-u` option on the bi-segment file. Thus, all the bi-segments are unique. However, there are cases where several different segments on one side have exactly the same translation. This means that segments on one side of the bi-text may be repeated. Only bi-segments are unique. This process was performed for each input chunk of text, not for the whole corpus.

2.4 Translation Models

All the training, tuning, and decoding was done with the Moses toolkit. Only results for hierarchical translation models are documented in this report. The hierarchical models (*I*) are based on Synchronous Context Free Grammars (SCFGs).

2.5 Training Outline

In this section, the steps of building the SMT system are briefly outlined.

2.5.1 Language Models

All LMs were trained with the `ngram-count` tool from the Stanford Research Institute Language Model (SRILM) toolkit (*12*).

All the LMs were of order 5, and the four LMs were trained separately. The LMs were named: artificial, legal, military, and news. The artificial model was trained on LDC phrasebook and elicitation data. The Legal model was trained on the Afghan constitution. The military model was trained on the data from FM 7-8. The news model was trained on the news text from the ARR, LDC, and SeA news corpora.

Each model became a weighted term in a log linear model. The individual weights for the LMs in the log linear model were tuned as feature function weights.

2.5.2 Word Alignment

Word alignments were obtained by a standard run of GIZA++ (9). The English-to-Foreign (E2F) and Foreign-to-English (F2E) directions of the alignments were symmetrized into a single alignment using the grow-diagonal-final heuristic.

2.5.3 Translation Tables

Lexical translation tables were produced. From that, the phrases and hierarchical SCFG rules were extracted. Finally, the rules were scored.

2.5.4 Tuning

MERT was run to tune the weights for a log linear model, combining the translation and language models.

2.6 Evaluation

I used the `multi-bleu.perl` tool distributed with the Moses toolkit to get the Bilingual Evaluation Understudy (BLEU) (10) scores comparing the decoder output and one reference. Each stage of the process was evaluated separately.

The text was separated into two groups: the experimental group containing the text from FM 7-8, and the control group from the Ranger Handbook.

3. Results

The Army FM 7-8 was translated in 13 stages.

3.1 Baseline on FM 7-8

Table 3 shows the BLEU scores that were obtained when the baseline system was run on each chunk of text from the FM 7-8.

Table 3. BLEU scores for the baseline system on each chunk of text from FM 7-8.

Chunk	BLEU Score
1	1.80
2a	2.19
2b	1.74
2c	1.65
2d	2.94
2e	2.45
2f	3.64
3a	3.54
3b	3.02
4	1.56
5a	2.04
5b	1.97
5c	2.11

3.2 Baseline on Ranger Handbook

Table 4 shows BLEU scores of the baseline system when run on the chapters of the ranger Handbook.

Table 4. BLEU scores for the baseline system on each chunk of text from the Ranger Handbook.

Chapter	BLEU Score
1	0.00
2	1.24
3	2.07
4	1.13
5	1.13
6	1.45
7	0.84
8	0.00
9	1.79
10	0.86
11	1.41
12	0.00
13	0.00
14	1.12

3.3 Stages

Consecutive chunks of text were translated at each stage. A hierarchical TM and 5-gram LM were retrained at each stage. The preceding chunks of text were used as training and tuning data at each stage. In general, scores increased as stages increased. Scores decreased in the following cases:

- Stage III on chunks 2d and 3a.
- Stage VI on chunk 3b and chapter 4.

The maximum score of 25.39 BLEU was achieved by the Stage X system on chunk 5a.

3.3.1 FM 7-8 on FM 7-8 by Stage

Table 5 shows the BLEU scores of the systems and test sets that were used to translate the FM 7-8 text.

Table 5. FM 7-8 BLEU scores for test of systems at each stage. The first column shows the text used in training and tuning. The second column indicates the test text. The third column shows the BLEU score for the test.

Training	Test	BLEU Score
base	1	1.80
base,1	2a	8.08
base,1,2a,	2b	6.16
base,1,2a,b	2c	6.44
base,1,2a,b,c	2d	8.48
base,1,2a,b,c,d	2e	8.20
base,1,2a,b,c,d,e	2f	8.61
base,1,2a,b,c,d,e,f	3a	13.88
base,1,2a,b,c,d,e,f,3a	3b	14.76
base,1,2a,b,c,d,e,f,3a,b	4	16.56
base,1,2a,b,c,d,e,f,3a,b,4	5a	25.39
base,1,2a,b,c,d,e,f,3a,b,4,5a	5b	23.03
base,1,2a,b,c,d,e,f,3a,b,4,5a,b	5c	18.57

3.3.2 Ranger Handbook on Ranger Handbook by Stage

Table 6 shows results for systems trained incrementally on chapters of the Ranger Handbook and tested on the immediately following chapter.

Table 6. Ranger Handbook BLEU scores for test of systems at each stage. The first column shows the text used in training and tuning. The second column indicates the text. The third column shows the BLEU score for the test.

Training Chapters	Test Chapter	BLEU Score
base	1	0.00
base,1	2	2.76
base,1,2	3	3.17
base,1,2,3	4	5.88
base,1,2,3,4	5	5.67
base,1-5	6	8.79
base,1-,6	7	1.79
base,1-7	8	4.88
base,1-8	9	3.67
base,1-9	10	3.03
base,1-10	11	2.64
base,1-11	12	2.74
base,1-12	13	2.61
base,1-13	14	3.42

3.3.3 Ranger Handbook on FM 7-8 Results by Stage

Table 7 shows BLEU scores for systems trained and tuned on the chapters from the Ranger Handbook and tested on the chunks of FM 7-8. Comparing this table with the analogous table of results for FM 7-8 trained and tune systems 5 illustrates the single translator over-fitting effect.

Table 7. BLEU scores for systems trained and tuned on Ranger Handbook text and tested on FM 7-8 at each stage. The first column shows the text used in training and tuning. The second column indicates the text. The third column shows the BLEU score for the test.

Training Chapters	Test Chunk	BLEU Score
base	1	1.80
base,1	2a	2.38
base,1,2	2b	3.03
base,1,2,3	2c	3.73
base,1,2,3,4	2d	3.68
base,1-5	2e	4.42
base,1-6	2f	3.70
base,1-7	3a	6.45
base,1-8	3b	5.93
base,1-9	chapter 4	8.20
base,1-10	5a	5.74
base,1-11	5b	3.73
base,1-12	5c	3.59

4. Conclusion

In this section, the main conclusion of the investigation is stated and a brief plan for future work is given.

4.1 Conclusions

The use of recent training and tuning data from the same text being translated yields faster improvements in accuracy than improvements observed on the same text by systems constructed with data from a different text with a similar domain and genre. A comparison of table B-1 with tables B-3 and B-4 in appendix B supports this conclusion. Improvements in accuracy are faster and larger when the method involves a single translator who reads the drafts produced by the SMT than when the translator does not get this feedback. A comparison of tables B-1 and B-2 in appendix B provide evidence that the project-specific method leads to models that are over-fit to the single translator.

4.2 Future Work

An effort will be undertaken to exploit the target input text to fit the system even further to the translation of the text. Given the input text I and a large set of in domain bi-segments $B = E \times F$ and a distance function $d : E \times I \rightarrow [0, \infty)$, where I represents the text to be translated and B

represents the set of available tuning data. Programs will be written that choose bi-segments $(e,f) \in B$ that are closest to the segments $j \in I$. That is, the tuning segments e will be chosen such that the distances $d(e,j)$ are minimal.

5. References

1. Chiang, D. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, June 2005, 263–270, Association for Computational Linguistics.
2. Hardt, D.; Elming, J. Incremental Re-training for Postediting SMT. *AMTA 2010*, November 2010.
3. Hoang, H.; Birch, A.; Callison-Burch, C.; Zens, R.; Constantin, A.; Federico, M.; Bertoldi, N.; Dyer, C.; Coean, B.; Shen, W.; Moran, C.; Bojar, O. *Moses: Open Source Toolkit for statistical Machine Translation*, 2007.
4. Institute for War and Peace. Afghan recovery report. <http://www.iwpr.net> (accessed 2011).
5. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, September 2005.
6. Levenberg, A.; Callison-Burch, C.; Osborne, M. Streambased Translation Models for Statistical Machine Translation. In *Proceedings of HLT-NAACL, 2010*, 394–402.
7. Linguistic Data Consortium. *Less Commonly Taught Languages, Pashto Language Pack v1.1*, 2007.
8. Och, F. J. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of ACL*, 2003.
9. Och, F. J.; Ney, H. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* **2003**, 29 (1), 19–51.
10. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 2002, 311–318.
11. Sada-e Azadi. <http://www.sada-e-azadi.net> (accessed 2011).
12. Stolcke, A. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*, Denver, CO, 2002.
13. U.S. Army Infantry School. *FM 7-8 Infantry Rifle Platoon and Squad*.
14. U.S. Army Infantry School. *Ranger Handbook*.

INTENTIONALLY LEFT BLANK.

Appendix A. Corpus Statistics

Table A-1 through A-7 include statistics about the corpora used in this project.

A-1. Numbers of Tokens in Baseline

Table A-1 shows the number of tokens from each corpus.

Table A-1. Number of tokens in each corpus.

Corpus	English Tokens	Pashto Tokens
ARR	52245	64782
LDC Elicitation	19705	18096
LDC News	214528	242137
LDC Phrasebook	8392	8041
Legal	10547	10578
QAMO dictionary	867	1556
SeA	305813	364811
total	612097	710001

A-2. Numbers of Segments from Chunks of FM 7-8

Table A-2 shows the number of bi-segments from each chunk of FM 7-8.

Table A-2. Number of bi-segments from chunks of FM 7-8.

Chunk	Segments
chapter 1	454
chapter 2 chunk a	482
chapter 2 chunk b	411
chapter 2 chunk c	600
chapter 2 chunk d	500
chapter 2 chunk e	356
chapter 2 chunk f	427
chapter 3 chunk a	360
chapter 3 chunk b	364
chapter 4	511
chapter 5 chunk a	328
chapter 5 chunk b	300
chapter 5 chunk c	300
total	5478

A-3. Total Number of Segments from FM 7-8

Table A-3 shows the cumulative number of segments from FM 7-8.

Table A-3. Cumulative number of segments from FM 7-8.

Stage	Total Segments
I	454
II	936
III	1347
IV	1947
V	2447
VI	2803
VII	3230
VIII	3593
IX	3957
X	4470
XI	4799
XII	5099
XIII	5399
XIV	5478

A-4. Numbers of Types

Table A-4 shows the total number of types in each training stage.

Table A-4. Total number of types in each training stage. Notice how slowly these numbers increase.

Stage	English	Pashto
baseline	22751	37528
I	23153	37645
II	23255	37801
III	23343	37927
IV	23453	38123
V	23564	38283
VI	23664	38572
VII	23779	38701
VIII	23850	38775
IX	23898	38866
X	23931	38967
XI	23961	39019
XII	23991	39059
XIII	24024	39154

A-5. Numbers of Tokens from FM 7-8

Here we list the number of tokens that occur in each chunk of text from FM 7-8.

Table A-5. Number of tokens in each chunk of FM 7-8.

Chunk	English Tokens	Pashto Tokens
1	7177	9477
2a	6147	7845
2b	5897	8031
2c	8183	11096
2d	7301	9383
2e	5254	6878
2f	6164	8063
3a	5638	7031
3b	5618	7224
4	8870	11088
5a	3326	4292
5b	3581	4478
5c	3795	4929
total	63381	94881

A-6. Cumulative Token Numbers

Here we list the total cumulative number of tokens from FM 7-8 that occur as the rebuilding process moves through the stages.

Table A-6. Cumulative number of tokens from FM 7-8 by stage.

Stage	Cumulative English Tokens	Cumulative Pashto Tokens
I	7177	9477
II	13324	17322
III	19221	25353
IV	27404	36449
V	34705	45832
VI	39959	52710
VII	46123	60773
VIII	51760	67794
IX	57373	75013
X	66248	86106
XI	69569	90393
XII	75425	94871
XIII	79220	99800

A-7. Ranger Handbook Segments

Table A-7 shows the numbers of bi-segments in each chapter of the Ranger Handbook.

Table A-7. Numbers of bi-segments in each chapter of the Ranger Handbook.

Chapter	Number of Segments
1	257
2	1192
3	226
4	301
5	935
6	482
7	141
8	126
9	345
10	234
11	545
12	248
13	93
14	343

Appendix B. System Scores

Tables B-1 through B-6 show more scores for the systems developed in this project. Systems are trained either on FM 7-8 or Ranger Handbook data.

B-1. All Scores

All tables show BLEU scores for all stages on all chunks of text. Note the higher scores that appear above the diagonal in these tables. The higher scores occur as a result of testing on the training data.

B-1.1 FM 7-8 on FM 7-8

Scores for systems built with FM 7-8 text and tested on FM 7-8 are shown in table B-1.

Table B-1. All FM 7-8 BLEU scores for systems by stage and FM 7-8 chunk.

Chunk	Base	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	1.80	69.84	68.74	67.07	67.36	66.18	66.62	68.27	66.58	66.73	65.76	66.63	65.04	65.84
2a	2.19	8.08	73.00	74.47	73.86	73.19	72.80	72.95	74.38	72.68	72.95	73.12	70.11	72.13
2b	1.74	5.37	6.16	67.48	68.23	68.45	68.87	68.52	69.53	69.37	69.09	69.29	68.21	66.24
2c	1.65	5.91	6.38	6.44	71.85	72.04	71.17	71.57	70.99	70.80	68.81	70.18	66.97	67.24
2d	2.94	5.83	7.37	6.96	8.48	77.01	76.36	77.21	77.23	77.17	77.09	75.17	75.75	76.00
2e	2.45	5.69	6.35	7.39	7.50	8.20	76.69	78.51	77.99	78.38	78.33	77.60	75.59	78.63
2f	3.64	4.08	5.43	5.91	6.67	8.38	8.61	79.95	80.30	81.25	81.40	80.41	78.83	80.92
3a	3.54	7.64	9.70	9.33	10.16	11.07	11.68	13.88	79.51	79.81	80.21	79.80	77.34	80.47
3b	3.02	6.91	7.57	8.63	9.50	11.26	11.01	12.32	14.76	80.79	80.06	79.91	78.76	81.51
4	1.56	7.47	9.03	10.23	11.62	13.25	12.78	13.86	15.62	16.56	77.61	76.77	75.45	77.84
5a	2.04	10.59	13.66	13.81	16.32	17.89	18.99	19.47	20.66	20.86	25.39	83.71	82.05	82.75
5b	1.97	8.39	9.06	11.81	11.82	14.48	14.43	15.73	17.22	18.62	19.34	23.03	81.25	84.59
5c	2.11	7.43	8.93	9.34	10.19	13.25	13.61	15.59	16.74	16.90	17.42	18.48	18.57	79.16

B-1.2 FM 7-8 on Ranger Handbook

BLEU scores for systems constructed with FM 7-8 data and tested on Ranger Handbook data are shown in table B-2.

Table B-2. All results of FM 7-8 system tested on Ranger Handbook.

Chapter	Base	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	0.00	2.27	2.61	2.63	2.74	2.83	2.84	2.54	2.97	3.00	3.26	3.62	3.12	4.67
2	1.24	2.49	3.72	4.16	4.19	4.14	4.61	4.44	4.24	4.13	4.06	4.34	3.94	5.18
3	2.07	1.37	2.98	2.62	2.37	2.66	2.60	2.56	2.34	3.39	2.70	2.58	2.23	3.19
4	1.13	3.19	3.75	3.79	3.69	4.29	4.02	4.35	4.11	4.13	4.40	4.41	4.34	4.14
5	1.13	2.50	2.90	2.93	3.46	3.56	3.75	3.78	4.75	4.77	4.75	4.80	4.60	5.47
6	1.45	3.71	3.75	3.72	4.02	4.59	4.07	4.25	4.35	4.48	8.71	8.37	8.13	8.77
7	0.84	1.20	1.03	1.16	1.13	1.45	0.00	0.00	0.91	0.99	2.01	1.50	0.78	1.64
8	0.00	2.93	2.83	3.20	3.44	4.57	4.00	4.51	3.44	3.56	4.56	3.75	3.33	4.77
9	1.79	1.69	1.91	2.00	1.75	2.13	1.80	1.50	1.99	2.14	1.67	1.86	1.56	2.03
10	0.86	1.11	1.12	1.22	1.56	1.57	1.80	1.18	1.29	1.22	1.49	1.17	1.29	1.76
11	1.41	1.72	1.75	1.77	1.92	2.12	2.14	1.98	2.27	2.36	2.34	2.09	1.87	2.70
12	0.00	1.09	0.81	0.00	0.00	0.92	0.98	0.00	1.75	1.28	1.41	2.09	1.53	2.00
13	0.00	1.34	1.61	1.66	1.49	1.49	1.72	1.37	1.31	1.56	2.04	1.40	1.52	1.63
14	1.12	1.51	1.91	1.91	2.14	2.20	2.78	2.41	2.63	2.50	3.05	3.09	2.38	3.20

B-1.3 Ranger Handbook on Ranger Handbook

BLEU scores for systems constructed with Ranger Handbook data and tested on the Ranger Handbook are shown in table B-3.

Table B-3. BLEU scores for Ranger Handbook trained and tuned systems tested on Ranger Handbook.

Chapter	Base	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	0.0	70.86	69.65	70.40	70.44	68.10	63.18	69.94	65.55	63.27	64.47	63.61	66.13	63.55
2	1.24	2.76	70.99	72.03	71.07	70.03	70.22	69.30	69.08	68.61	69.21	68.17	68.39	67.17
3	2.07	1.21	3.17	76.46	71.76	75.49	74.64	72.14	74.15	69.29	71.79	75.47	72.07	71.26
4	1.13	2.61	5.15	5.88	73.22	70.45	71.63	70.94	70.92	70.29	69.87	70.12	72.06	67.81
5	1.13	2.61	5.56	5.38	5.67	67.77	69.08	68.37	67.36	66.80	66.26	66.48	67.27	66.01
6	1.45	3.15	5.00	5.14	5.10	8.79	69.75	67.44	67.76	67.83	67.97	66.17	66.09	65.99
7	0.84	0.75	1.54	0.00	0.00	1.97	1.79	43.99	42.76	41.97	40.25	38.25	39.27	37.91
8	0.0	2.28	3.55	2.83	3.78	5.73	5.77	4.88	58.06	56.76	60.46	57.51	53.33	55.43
9	1.79	1.71	2.85	2.77	2.89	4.11	3.98	4.56	3.67	63.25	63.13	61.65	63.92	61.38
10	0.86	0.00	1.97	1.31	0.78	1.59	1.74	2.23	1.49	3.03	59.86	59.84	61.92	59.24
11	1.41	1.59	2.03	2.28	1.95	2.87	2.68	2.82	2.90	3.25	2.64	62.18	66.09	61.77
12	0.0	0.83	1.83	1.45	1.40	2.00	2.40	1.83	3.24	2.68	2.03	2.74	67.61	65.04
13	0.0	1.31	1.60	1.53	1.41	1.45	1.10	1.17	1.21	2.06	2.24	2.10	2.61	57.47
14	1.12	1.70	2.81	3.03	2.98	3.68	4.45	4.26	3.90	4.40	3.75	3.88	4.26	3.42

B-1.4 Ranger Handbook on FM 7-8

BLEU scores for systems constructed with Ranger Handbook data and tested on FM 7-8 are shown in table B-4. Recall that these systems were not developed with a translator who received feedback from systems in previous iterations.

Table B-4. BLEU scores for tests of systems trained and tuned on Ranger Handbook data and evaluated on FM 7-8.

Chunk	Base	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	1.80	1.96	3.31	3.33	3.86	4.88	5.40	5.12	5.50	4.90	4.65	4.30	4.87	4.36
2a	2.19	2.38	5.45	5.56	5.68	5.56	5.29	5.16	5.41	5.49	4.53	4.69	5.02	4.64
2b	1.74	1.72	3.03	3.10	3.08	3.48	3.49	3.42	3.20	3.09	2.89	2.50	2.84	2.72
2c	1.65	2.03	3.40	3.73	2.98	3.50	3.94	3.74	3.39	3.79	3.13	3.10	3.01	2.54
2d	2.94	2.40	4.16	4.19	3.68	4.38	4.32	4.44	4.73	4.44	3.97	3.22	4.07	3.38
2e	2.45	2.50	4.27	4.25	4.06	4.42	4.60	4.90	4.69	4.04	3.51	3.65	4.41	3.19
2f	3.64	2.25	3.29	2.96	2.42	3.04	3.70	3.35	3.35	3.37	3.10	2.80	3.20	2.52
3a	3.54	3.29	4.53	4.75	4.48	6.41	6.91	6.45	7.08	6.59	6.54	6.12	6.31	5.71
3b	3.02	1.82	4.29	4.37	3.38	5.81	5.52	5.82	5.93	5.91	5.36	4.81	5.45	4.94
4	1.56	2.83	4.90	5.09	4.49	4.86	7.69	7.69	7.94	8.20	7.84	7.79	7.80	7.25
5a	2.04	3.64	5.59	5.94	5.59	5.04	5.35	5.12	6.10	6.35	5.74	5.21	5.31	5.46
5b	1.97	2.49	4.95	4.49	4.20	4.06	4.53	4.20	4.69	4.37	5.06	3.73	4.68	3.86
5c	2.11	2.75	4.23	4.17	3.76	4.50	4.23	4.26	4.30	4.39	3.90	3.79	3.59	3.38

B-2. Tuning versus No Tuning

In this appendix, tuned and non-tuned systems are compared so see if tuning makes a difference.

B-2.1 Stage IX No Tuning

BLEU scores comparing tuned and non-tuned systems for stage IX are shown in table B-5. The systems were trained and tuned on the Ranger handbook. The tests are on the chunks from FM 7-8.

Table B-5. BLEU scores for two systems at stage IX: one with tuning and one without tuning.

Chunk	IX	IX Plus Tuning
FM 7-8 1	3.64	4.90
FM 7-8 2a	3.94	5.49
FM 7-8 2b	3.43	3.09
FM 7-8 2c	3.03	3.79
FM 7-8 2d	4.38	4.44
FM 7-8 2e	3.04	4.04
FM 7-8 2f	4.51	3.37
FM 7-8 3a	5.72	6.59
FM 7-8 3b	5.04	5.91
FM 7-8 4	4.32	8.20
FM 7-8 5a	4.91	6.35
FM 7-8 5b	4.79	4.37
FM 7-8 5c	4.32	4.39
Ranger Handbook (RH) 1	11.36	63.27
RH 2	16.29	68.61
RH 3	19.78	69.29
RH 4	8.65	70.29
RH 5	10.35	66.80
rh 6	10.85	67.83
rh 7	10.92	41.97
rh 8	9.46	56.76
RH 9	10.65	63.25
RH 10	1.33	3.03
RH 11	1.78	3.25
RH 12	1.36	2.68
RH 13	0.0	2.06
RH 14	2.33	4.40

B-2.2 FM 7-8 Stage X No Tuning

BLEU scores comparing tuned and non-tuned systems for stage X are shown in table B-6. The systems were trained on and tested on chunks of FM 7-8.

Table B-6. BLEU scores for two stage-X systems: one with tuning and one without tuning.

Chunk	Tuned	No Tuning
1	65.76	10.97
2a	72.95	14.13
2b	69.09	11.12
2c	68.81	12.01
2d	77.09	17.09
2e	78.33	17.03
2f	81.40	19.39
3a	80.21	19.02
3b	80.06	17.21
4	77.61	19.17
5a	25.39	11.51
5b	19.34	10.87
5c	17.42	9.57
RH 1	3.26	1.72
RH 2	4.06	2.99
RH 3	2.70	1.73
RH 4	4.41	2.18
RH 5	4.75	2.34
RH 6	8.71	2.50
RH 7	2.01	1.19
RH 8	4.56	3.23
RH 9	1.67	2.24
RH 10	1.49	1.34
RH 11	2.34	1.38
RH 12	1.41	0.00
RH 13	2.04	0.00
RH 14	3.05	1.81

List of Symbols, Abbreviations, and Acronyms

ANA	Afghan National Army
ARL	U.S. Army Research Laboratory
ARR	Afghan Recovery Report
BLEU	Bilingual Evaluation Understudy
E2F	English-to-Foreign
EM	Expectation Maximization
F2E	Foreign-to-English
FM	Field Manual
LCTL	Less Commonly Taught Languages
LDC	Linguistic Data Consortium
LM	language model
MERT	Minimum Error Rate Training
MLCB	Multilingual Computing Branch
MT	machine translation
NVTC	National Virtual Translation Center
QAMO	Qamoosuna (dictionaries)
RH	Ranger Handbook
SCFG	Synchronous Context Free Grammar
SeA	Sada-e-Azadi
SME	subject matter expert
SMT	Statistical Machine Translation
SRILM	Stanford Research Institute Language Model
TM	translation model

NO. OF COPIES	ORGANIZATION
1 ELEC	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1	US ARMY RSRCH DEV AND ENGRG CMND ARMAMENT RSRCH DEV & ENGRG CTR ARMAMENT ENGRG & TECHNLOGY CTR ATTN AMSRD AAR AEF T J MATTS BLDG 305 ABERDEEN PROVING GROUND MD 21005-5001
1	US ARMY INFO SYS ENGRG CMND ATTN AMSEL IE TD A RIVERA FT HUACHUCA AZ 85613-5300
1	COMMANDER US ARMY RDECOM ATTN AMSRD AMR W C MCCORKLE 5400 FOWLER RD REDSTONE ARSENAL AL 35898-5000
1	US GOVERNMENT PRINT OFF DEPOSITORY RECEIVING SECTION ATTN MAIL STOP IDAD J TATE 732 NORTH CAPITOL ST NW WASHINGTON DC 20402
17 HCS 1 ELEC	US ARMY RSRCH LAB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ATTN RDRL CII B BROOME ATTN RDRL CII T J J MORGAN (12 HCS, 1 PDF) ATTN RDRL CII T V M HOLLAND ATTN RDRL CIO LL TECHL LIB ATTN RDRL CIO MT TECHL PUB ADELPHI MD 20783-1197

TOTAL: 24 (2 ELEC, 1 CD, 21 HCS)