ARMY RESEARCH LABORATORY

Improving Domain-specific Machine Translation by Constraining the Language Model

by Jeffrey C. Micher

ARL-TN-0492 **July 2012**

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-TN-0492 **July 2012**

Improving Domain-specific Machine Translation by Constraining the Language Model

Jeffrey C. Micher
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)
July 2012	Final	

4. TITLE AND SUBTITLE
Improving Domain-specific Machine Translation by Constraining the Language Model

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jeffrey C. Micher

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-TN-0492

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
A domain-specific statistical machine translation engine is shown to be more accurate when only domain-specific language data are used to build the target-language language model. This has been found to be true when compared to using a much larger, out-of-domain corpus for building the language model, either alone or in combination with the domain-specific data.

15. SUBJECT TERMS
Domain-specific machine translation customization language model

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jeffrey C. Micher
Unclassified	Unclassified	Unclassified	UU	18	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-0316

Contents

List of Tables

1. Introduction

General purpose machine translation (MT) engines have improved dramatically over the last two decades. However, when translating material that is specific to a particular domain, general-purpose engines often perform poorly. To address this problem, various means of customization have been proposed. One such means involves creating domain-specific statistical MT systems, but there are many ways this can be accomplished. Here, we explore the use of an in-domain language model versus a general-domain, larger language model in conjunction with a domain-specific translation model in a statistical MT system to improve translation of domain-specific text.

2. Background

Statistical MT systems make use of parallel corpora to estimate the probabilities of word and phrase translations, and the probabilities of how these are put together to make sentences. From a very simplified point of view, they do this with two main components, a translation model, which provides the most likely translations of source words and phrases, and a target language model, which helps to identify the most likely sequence of these translated pieces.

With these two main components in mind, in statistical MT, it is generally assumed that the use of more training data will produce better results. More examples of translations should mean (1) better estimations of the probabilities of those translations and (2) better translation coverage, resulting in better MT. The field of statistical MT has held this notion as fundamental and has always advocated the improvement of MT systems first and foremost through the use of greater amounts of training data in the two models, especially in the target language model (Brants et al., 2007). Och (2005) reports findings of using varying amounts of target language training data, which show incremental system performance with greater and greater amounts of data. At the National Institute of Standards and Technology (NIST) ’06 Machine Translation Evaluations, the highest scoring systems were those that were able to train with the largest language models (NIST, 2006). The highest scoring Arabic-English system used a 1-trillion-word language model (Och, 2006). The next highest scoring system used 33 million words in the language model (Chiang et al., 2006).

However, narrow domains generally do not have much training data available, so it is impossible to create a system with a very large corpus of domain-specific training data to improve its performance. To make up for the lack of parallel training data, one assumption is that more monolingual target language data should be used in building the target language model. Prior work on domain-specific MT has focused on training target language models with monolingual

domain-specific data. Eck et al., (2004a) show a significant improvement in performance on a Chinese-English statistical system when a language model is built using an information retrieval technique. Sentences relevant to the test document are retrieved from a corpus and used to build the language model. Xu et al. (2007) use domain-specific language models with an engine trained on general data and show improvements over using a general language model. In both of these papers, however, the translation model training data are large and not domain-specific.

Here we propose a novel approach, which uses a small amount of domain-specific parallel training data along with a target language model also trained with a small amount of domain-specific data. We show that this configuration improves performance over systems whose language model is trained with larger amounts of out-of-domain data, even when the size of the parallel data is small.

In a previously unpublished study with narrow domain MT (a graduate student project [Micher, 2003]), it was revealed that the use of a large corpus of out-of-domain, more general data does not necessarily improve an MT system that is targeted at translating in a narrow domain. The MT system used was an example-based machine translation (EBMT) system from Carnegie Mellon University, PanEBMT (Brown, 1996). For this project, 6.7k lines of parallel French/English text from a computer manual (Semantic Compaction Systems and Prentke Romich Company) along with 100k lines of the Canadian Hansards (UPenn, 2010a) French/English parallel corpus were used in the experiment. In this report, the Hansards corpus is referred to as "H" and the computer manuals referred to as "D" for "domain-specific." A test set was created from the D corpus by holding out 100 sentence pairs by systematic selection: every 67th sentence pair in the corpus. Bilingual Evaluation Understudy (BLEU)-4 (Papineni et al., 2002) was used to evaluate the MT results, using one reference translation.

The experiment was set up as follows. Three EBMT systems were created: (1) using the H corpus alone, (2) using the H+D corpora, and (3) using the D corpus alone. The results of the experiment are summarized in table 1.

Table 1. BLEU scores on EMBT system.

Training Set:	**H**	**H+D**	**D**
BLEU-4	15.52	27.96	27.90

As can be seen, there is an expected increase in system performance when adding domain-specific data to the training data for the system. However, when removing the larger, out-of-domain data from the training set, leaving just the in-domain data, an unexpected stability in system performance is observed. The system trained with only domain data does no worse than the system trained with the larger data set. These results suggest that building an MT system with a large amount of more general, generally unrelated data do not necessarily improve an MT system.

Carrying this idea further, it is hypothesized that a statistical MT engine built with domain-specific data for both the translation model and the language model should perform similarly to the EBMT system presented above.

3. Experimental Design

3.1 Data

For the current experiment, 38,970 lines of parallel Arabic-English military training data were used, consisting of approximately 500k tokens in each language. The corpus was automatically extracted from training manuals and materials. It was then hand-aligned by a native speaker who was also a military subject matter expert. Since the data contained substantial outline formatting (numbers and letters followed by periods and/or parentheses), these format indicators were removed automatically. The data also had a number of broken hyphenations, which remained after the automatic extraction process. These were fixed automatically. Spot checking revealed additional areas where the text was misaligned, so these areas were hand corrected.

The Arabic data were then transcribed automatically from Arabic script to Buckwalter (Liberman) encoding and morphologically analyzed. The best analysis was selected using ARAGEN (Habash, 2004), a morphological analyzer that is built on top of the analysis algorithm from the Buckwalter Morphological Analyzer (UPenn, 2010b). Then, both the English and Arabic text was tokenized to separate punctuation from words.

The English section of the European Parliamentary Proceedings corpus (Europarl corpus [Koehn, 2005]) was used to build the more general, out-of-domain, larger, target language model. This corpus contained 1,334,094 lines of text, consisting of 36,436,449 tokens and 98,954 individual types. A comparison of the sizes of the corpora that were used is summarized in table 2.

Table 2. Corpus sizes compared.

	Lines	**Tokens**	**Types**
Military Training Materials	38,970	508,985	16,430
Europarl Corpus	1,334,094	36,436,449	98,954
Ratio Military to Europarl	≅ 1/34	≅ 1/71	≅ 1/6

3.2 Experiment

The experiment was set up as follows. Five training and testing sets were created by randomly sampling 500 parallel lines from the military data for testing and leaving the remaining data for training. Five separate systems were then created with the military training data sets, using the

Moses statistical MT software (Moses, 2012). For each system, three language models were created: one using only the English side of the military data, one that combined both the military data and the English part of the Europarl corpus, and one that was built with only the English Europarl corpus. Each of the five systems was then tested by translating its respective test set three times, using the three different language models, but with the translation models trained on only the smaller military data. For each test result, Bleu-4 scores were calculated using one reference translation and are recorded in the table 3.

Table 3. Experimental results.

Build	Military Only LM	Military + Europarl LM	Europarl Only LM
1	0.2453	0.2351	0.1445
2	0.2421	0.2347	0.1420
3	0.2468	0.2351	0.1383
4	0.2401	0.2322	0.1411
5	0.2392	0.2292	0.1368

As can be seen from the data, all builds show that there is an increase in the BLEU score when using the language model built from adding the domain-specific data to the Europarl corpus. There is also an increase when removing the Europarl data from the language model. The systems using only domain-specific data for the language model scored the highest. These data show the same pattern as with EBMT builds; however, in this experiment, there are even slightly better scores using the domain-specific language model alone.

4. Discussion

These data certainly seem to contradict the belief that more data means better translations. One of the reasons for this divergence is that systems built from general or out-or-domain data lack domain-specific key terminology. In fact, addition of domain terminology has been shown to improve performance of generalized MT systems. Eck et al. (2004) showed that the using a large dictionary extracted from medical domain documents in a statistical MT system to generalize the training data significantly improves the translation performance.

Comparison of the 1-, 2-, and 3-grams from the two training corpora in this study suggests that there is a lack of domain-specific terminology in the Europarl data (table 4). Only 12.39% of the unigrams from the military corpus are repeated in the Europarl corpus, and as the n-gram size increases, the percentage of overlap gets dramatically smaller.

Table 4. N-grams in corpora compared.

	Military	Europarl	M ∩ E	% of M in E
Unigram Types	16,488	98,954	12,263	12.39
Bigram Types	149,976	2,359,424	82,597	3.5
Trigram Types	288,825	10,163,466	81,967	0.8

Looking at frequency counts for each corpus, it's possible to see how military terminology is more prevalent in the military corpus than in the Europarl corpus. The lists in appendix A show the 10 most frequent 1-, 2-, and 3-grams overlapping in the corpora, but sorted by the ratio of occurrences in the military corpus compared to the Europarl corpus. For example, a domain-specific word "platoon" occurs much more frequently in the military corpus than in the Europarl corpus. This ratio is expected to be higher than a very frequent word in both corpora, such as "the." The ratio for "platoon" is 2108 (instances in military corpus) divided by 2 (instances in Europarl corpus) = 1054. The ratio for "the" is 0.02, and most of the most frequent words in both corpora have ratios less than 1. Thus, it is easy to see that "military" words in the military corpus are more frequent than in the Europarl corpus.

But what is it about removing the larger Europarl corpus from the training data that produces an increase in the BLEU score when translating military data? An explanation for this may be found by looking at lexical items that are ambiguous with respect to their target language translations. The larger, more general language model may have more instances of out-of-domain translations and prefer these when given a choice between in- or out-of-domain translations. This creates a "muddying" effect in the data when using the larger language model. When these general translations for specific domain vocabulary are removed from the language model training data, the domain-specific translations have a greater probability for given domain-specific terminology. To demonstrate this, in appendix B, we show five domain-specific words in Arabic, which could be used in a general sense, examining the probabilities for the translations that are given in the three language models. Three of these lexical items support this hypothesis, whereas only two support the idea that adding domain terminology to the language model improves its chances of getting selected. With all of these words, though, in the military-only language model, the military translation is the most probable out of the possible translations. Probabilities are given as log probabilities, so the closer the negative number is to zero, the higher its probability.

5. Conclusion and Future Work

We have shown that using a domain-specific language model in a statistical MT system produces better translations, even when that language model is smaller than a larger out-of-domain language model. We have looked at why this is by looking at frequency counts of 1-, 2-, and 3-grams that appear in both corpora. We have examined probabilities of domain-specific versus generic translations of ambiguous domain terminology and have postulated some explanations for the higher BLEU scores when removing the larger, out-of-domain data from the language model training set.

We used the Europarl corpus in this study because it was readily available and large. One could argue that the Europarl corpus itself is domain-specific, even though it is very large. Therefore, future work should include using other large corpora. It will also be important to devise an empirical definition of "domain" so that comparisons of corpora can be made with respect to domain specificity.

6. References

Brants, T.; Popat, A. C.; Xu, P.; Och, F. J.; Dean, J. Large Language Models in Machine Translation. Joint Meeting of the Conference on Empirical Methods on Natural Language *Processing and Conference on Computational Natural Language Learning following ACL* 2007.

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; Mercer, R. L. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 1993.

Brown, R. D. Example-Based Machine Translation in the Pangloss System. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96),* Copenhagen, Denmark, August 5-9, 1996, 169–174.

Chiang, D.; DeNeefe, S.; Fraser, A.; Graehl, J.; Hermjakob, U.; Knight, K.; Marcu, D.; Munteanu, D. S.; May, J.; Pust, M.; Soricut, R.; Voeckler, J. ISI at NIST'06, *NIST Machine Translation Workshop,* September 6–7, 2006.

Eck, M.; Vogel, S.; Waibel, A. Language Model Adaptation for Statistical Machine Translation based on Information Retrieval. *Proceedings of LREC,* 2004a.

Eck, M.; Vogel, S.; Waibel, A. Improving Statistical Machine Translation in the Medical Domain using the Unified Medical Language System. In Proceedings of *Coling 2004*, Geneva, Switzerland, August 2004b.

Habash, N. Large Scale Lexeme Based Arabic Morphological Generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04).* Fez, Morocco, 2004.

Koehn, P. *Europarl: A Parallel Corpus for Statistical Machine Translation*, MT Summit 2005. http://www.statmt.org/europarl/ (accessed June 2012).

Liberman, M. "Arabic Transliteration/Encoding Chart," http://www.ldc.upenn.edu/myl/morph/buckwalter.html (accessed June 2012).

Micher, J. Carnegie Mellon University, Machine Translation Lab, 11-732, final project, 2003.

Moses open source software, 24 April 2012. http://www.statmt.org/moses (accessed June 2012).

NIST, "NIST 2006 Machine Translation Evaluation Official Results," 1 November 2006. http://www.itl.nist.gov/iad/mig//tests/mt/2006/doc/mt06eval_official_results.html (accessed June 2012).

Och, F. J. Tutorial at MT Summit 2005, Phuket, Thailand, 2005.

Och, Franz J. The Google Statistical Machine Translation System for the 2006 NIST MT Evaluation, *NIST Machine Translation Workshop*, September 6–7, 2006.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, 2002, 311–318.

Semantic Compaction Systems and Prentke Romich Company, 2004.

University of Pennsylvania, Linguistic Data Consortium. Buckwalter Arabic Morphological Analyzer Version 2.0, 2010a. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004L02 (accessed June 2012).

University of Pennsylvania, Linguistic Data Consortium. "Hansard French/English," 2010b. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20 (accessed June 2012).

Xu, J.; Deng, Y.; Gao, Y.; Ney, H. Domain Dependent Statistical Machine Translation. *Proceedings of the MT Summit XI*, Copenhagen, Denmark, 10–14 September 2007.

Appendix A. Most Frequent N-grams based on Ratio between Corpora

Table A-1 shows the most frequent n-grams based on the ratio between the corpora.

Table A-1. Most frequent n-grams bsaed on the ratio between the corpora.

unigram	*Mil*	*Europarl*	*Ratio*
platoon	2108	2	1054.00
commanders	1522	16	95.13
squad	884	21	42.10
slide	1607	58	27.71
commander	2037	76	26.80
captive	806	41	19.66
enemy	2309	322	7.17
command	1478	253	5.84
fire	1260	773	1.63
units	876	551	1.59
bigram			
the captive	615	1	615.00
the commander	928	5	185.60
's intent	305	2	152.50
army leaders	245	3	81.67
the casualty	306	5	61.20
command and	326	7	46.57
of command	328	29	11.31
the enemy	1282	135	9.50
(see	347	37	9.38
of operations	313	87	3.60
trigram			
concept of operations	146	1	146.00
the enemy .	173	2	86.50
of the enemy	145	2	72.50
in this unit	71	1	71.00
of the commander	65	1	65.00
on the enemy	63	1	63.00
the enemy and	62	1	62.00
avenues of approach	55	1	55.00
command and control	187	4	46.75
of command and	66	2	33.00

INTENTIONALLY LEFT BLANK.

Appendix B. Log Probabilities of Selected Translations

Table B-1 shows the log probabilities of selected translations.

Table B-1. Log probabilities of selected translations.

Arabic Word:	أمر	**>mr**	
Translations	E only	E+M	M only
order	-3.17	-3.82	**-3.23**
matter	-3.32	-3.82	-4.25
issue	**-3.13**	**-3.36**	-3.57
Arabic Word:	مهمة	**mhmp**	
Translations	E only	E+M	M only
task	-3.85	-3.74	-3.15
assignment	-5.86	-4.86	-4.16
mission	-4.33	-3.77	**-2.92**
important	**-2.91**	**-3.58**	-3.48
serious	-3.53	-3.65	-3.97
Arabic Word:	الاستطلاع	**AlAstTlAE**	
Translations	E only	E+M	M only
reconnaissance	-6.51	-4.38	**-3.03**
poll	-5.63	-5.12	n/a
investigation	**-4.43**	**-4.03**	-4.53
Arabic Word:	سرية	**sryp**	
Translations	E only	E+M	M only
squadron	-7.28	-5.97	-5.05
secret	-4.54	-4.07	-4.52
private	**-3.93**	-3.75	-4.44
company	-4.08	**-3.63**	**-3.17**
Arabic Word:	مشاة	**m$Ap**	
Translations	E only	E+M	M only
infantry	-6.58	**-4.54**	**-3.37**
pedestrians	**-5.75**	-5.16	n/a

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	US ARMY INFO SYS ENGRG CMND ATTN AMSEL IE TD A RIVERA FT HUACHUCA AZ 85613-5300
15	US ARMY RSRCH LAB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ATTN RDRL CII B BROOME ATTN RDRL CII T C VOSS ATTN RDRL CII T J MICHER (5 HCS) ATTN RDRL CII T D BRIESCH ATTN RDRL CII T L HERNANDEZ ATTN RDRL CII T R HOBBS ATTN RDRL CII T S LAROCCA ATTN RDRL CII T V M HOLLAND ATTN RDRL CIO LL TECHL LIB ATTN RDRL CIO LT TECHL PUB ADELPHI MD 20783-1197