AFRL-RH-WP-TR-2013-0139

IMPROVED PHRASE TRANSLATION MODELING USING MAXIMUM A-POSTERIORI (MAP) ADAPTATION

Timothy Anderson, Ph.D.
Human Trust and Interaction Branch
Human-Centered ISR Division
Wright-Patterson AFB, OH 45433

A. Ryan Aminzadeh
US Department of Defense

Jennifer Drexler/Wade Shen
Massachusetts Institute of Technology Lincoln Laboratory

JULY 2013
Interim Report

AIR FORCE RESEARCH
LABORATORY 711TH HUMAN
PERFORMANCE WING HUMAN
EFFECTIVENESS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2013-0139 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//

TIMOTHY ANDERSON, Ph.D.
Work Unit Manager
Human Trust and Interaction Branch

//signature//

LOUISE CARTER, Ph.D.
Chief, Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

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 this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this 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)* 01-JUL-2013	**2. REPORT TYPE** Interim	**3. DATES COVERED** *(From - To)* 01 Jul 11 – 30 Jun 12

4. TITLE AND SUBTITLE Improved Phrase Translation Modeling using Maximum A-Posteriori (MAP) Adaptation	**5a. CONTRACT NUMBER** N/A
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Timothy Anderson, PhD. A. Ryan Aminzadeh Jennifer Drexler Wade Shen	**5d. PROJECT NUMBER** H0A3
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER H0A3 (5324X01S)
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES) Human Trust and Interaction Branch Human-Centered ISR Division 7llth Human Performance Wing Human Effectiveness Directorate Wright-Patterson AFB, OH 45433	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate Human-Centered ISR Division Human Trust and Interaction Branch Wright-Patterson AFB OH 45433	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S):** AFRL-RH-TR-WP-2013-0139

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A. Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

88ABW-2014-2275; Cleared 14 May 2014

14. ABSTRACT

In this paper, we explore several methods of improving the estimation of translation model probabilities for phrase-based statistical machine translation given in-domain data sparsity. We introduce a hierarchical variant of MAP adaptation for domain adaptation with an arbitrary number of out-of-domain models. We compare this adaptation technique to linear interpolation and phrase table fill-up. Additionally, we note that domain adaptation can have a smoothing effect, and we explore the interaction between smoothing and the incorporation of out-of-domain data. We find that the relative contributions of smoothing and interpolation depend on the datasets used. For both the IWSLT 2011 and WMT 2011 English-French datasets, the MAP adaptation method we present improves on a baseline system by 1.5+ BLEU points.

15. SUBJECT TERMS

Phrase Based Machine Translation, Domain Adaptation, Statistical Machine Translation, SMT, Maximum A-Posteriori Adaptation, MAP Adaptation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Timothy Anderson, Ph.D
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	15	**19b. TELEPHONE NUMBER** *(include area code)*

THIS PAGE INTENTIONALLY LEFT BLANK.

TABLE OF CONTENTS

Section **Page**

LIST OF FIGURES

Figure **Page**

LIST OF TABLE

Table **Page**

IMPROVED PHRASE TRANSLATION MODELING USING MAP ADAPTATION

A. Ryan Aminzadeh[1], Jennifer Drexler[2], Timothy Anderson[3], and Wade Shen[2]*

[1] US Department of Defense
ryan.aminzadeh@gmail.com
[2] Massachusetts Institute of Technology, Lincoln Laboratory
{j.drexler,swade}@ll.mit.edu
[3] Air Force Research Laboratory
timothy.anderson@wpafb.af.mil

1.0 ABSTRACT

In this paper, we explore several methods of improving the estimation of translation model probabilities for phrase-based statistical machine translation given in-domain data sparsity. We introduce a hierarchical variant of MAP adaptation for domain adaptation with an arbitrary number of out-of-domain models. We compare this adaptation technique to linear interpolation and phrase table fill-up. Additionally, we note that domain adaptation can have a smoothing effect, and we explore the interaction between smoothing and the incorporation of out-of-domain data. We find that the relative contributions of smoothing and interpolation depend on the datasets used. For both the IWSLT 2011 and WMT 2011 English-French datasets, the MAP adaptation method we present improves on a baseline system by 1.5+ BLEU points.

2 INTRODUCTION

Real-world performance of statistical MT models is often limited by training bitext availability. Performance of SMT models is sensitive not only to the amount of training data, but also to the domain from which these data are drawn. For optimal performance, developers of SMT systems will typically make significant investments to acquire bitexts *in the domain of interest*, which can be difficult and expensive.

In this paper, we describe efforts to improve SMT performance through better use of out-of-domain bitexts. We do this in the context of MAP estimation, a well-known approach for adaptation of statistical models that has been applied to the related problems of speech recognition [1] and language modeling [2,3], and which we apply here to the phrase translation tables used during SMT. We extend this method, which we developed previously for two phrase tables [4], to an arbitrary number of models.

*This work is sponsored by the Air Force Research Laboratory under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

In MAP adaptation, translation probabilities from the in-domain model are backed-off to out-of-domain estimates when the phrase pair occurs rarely in the in-domain data. When no out-of-domain estimate exists, this results in a smoothing effect. Linear interpolation, another domain adaptation technique, also has a smoothing effect for phrase pairs that are not contained in all of the models being interpolated.

When applied to a single phrase table, smoothing has been shown to improve SMT performance [5]. This observation leads us to investigate the extent to which the gains produced by domain adaptation are a result of smoothing.

In this paper, we empirically explore the relationship between phrase table smoothing and interpolation methods. We compare the results of MAP adaptation with the results of two other domain adaptation methods, linear interpolation and phrase table fill-up, and test these techniques on both smoothed and unsmoothed phrase tables.

3.1 Prior Work

3.2 Domain Adaptation

Maximum a-posteriori (MAP) adaptation is a Bayesian estimation method that attempts to maximize the posterior probability of a model given the data, as in [1]. The standard MAP formulation defines two probability distributions: a prior distribution ($p(s|t,\ \lambda 0$)) and a distribution estimated over the adaptation data ($p(s|t,\ \lambda adapt$). For phrase table estimation, the in-domain corpus is treated as adaptation data for estimating $p(s|t,\ \lambda adapt$) and the out-of-domain corpus is assumed to be $p(s|t,\ \lambda 0$).

$$\hat{p}(s|t,\lambda) = \frac{N_{adapt}(s,t)}{N_{adapt}(s,t)+\tau} p(s|t,\lambda_{adapt}) + \frac{\tau}{N_{adapt}(s,t)+\tau} p(s|t,\lambda_0) \qquad (1)$$

where $Nadapt$ ($s,\ t$) is the joint count of s and t in the adaptation data and τ is the MAP relevance factor. We previously used this method to effectively make use of out-of-domain data to improve performance of phrase tables trained with limited amounts of data[4]. Foster et al. [6] also used a MAP-based approach for phrase table interpolation, improving performance over a baseline system for a number of different datasets.

Linear interpolation of translation probabilities, with fixed weights for each corpus, has also been shown to improve SMT performance [7]. The basic formulation is:

$$\hat{p}(s|t,\lambda) = \sum_{i=1}^{M} \alpha_i(s,t) p(s|t,\lambda_i) \qquad (2)$$

where M is the number of corpora and αi is the interpolation coefficient for the ith corpus. Foster and Kuhn [7] test several strategies for determining the best mixture weights, comparing uniform weights with TF/IDF-, perplexity-, and EM- based techniques. Although they obtain small improvements using the more complex techniques, all variations yield performance gains of about 1 BLEU point on the NIST 2006 Chinese dataset.

Phrase table fill-up has also recently been used for domain adaptation [8]. Fill-up is initialized with all phrase pairs from the in-domain phrase table. Phrase pairs from the out- of-domain table are then added only if they are not in the in-domain table. The probabilities associated with phrase pairs added to the table are not changed. An extra binary feature is added to the table, indicating which model each phrase came from; that feature can then be used during optimization to penalize out-of-domain phrase pairs relative to those from the in-domain table. Additional phrase tables can be "cascaded" together, with phrase pairs from a new table added only when they are not already contained in the filled-up table. Another binary feature is added to the final phrase table for each additional model. On the 2011 IWSLT English-French dataset, fill-up adaptation improves performance by 0.7 BLEU points, and is comparable to linear interpolation with uniform weights.

3.3 Phrase Table Smoothing

Foster et al. [5] review a large number of phrase table smoothing techniques, and find that all of them significantly improve the performance of a baseline SMT system. Kneser-Ney [9] and modified Kneser-Ney [10] smoothing have the best performance overall, producing gains of almost 1.5 BLEU points on an English-French task. The phrase table formulation of KN smoothing is [5]:

$$p(s|t) = \frac{N(s,t) - D}{N(t)} + \alpha(t)P_b(s) \tag{3}$$

$$\alpha(t) = Dn_{1+}(*,t)/\sum_{\tilde{s}} N(\tilde{s},t) \tag{4}$$

$$P_b(s) = n_{1+}(s,*)/\sum_{\tilde{s}} n_{1+}(\tilde{s},t) \tag{5}$$

where $n1+(s, *)$ is the number of target phrases aligned with source phrase s, and $n1+(*, t)$ is the reverse. In modified Kneser-Ney, the discount D is replaced by an empirically derived discount Di, which is dependent on the joint count $N(s, t) = i$ of the source-target phrase pair.

Chen et al. [11] introduce an enhanced low frequency (ELF) feature designed to penalize phrase pairs with low joint counts during optimization. The feature is: $h_{elf}(s, t) = e^{(-1/N(s,t))}$, which is a $1/N(s, t)$ penalty in a log-linear model. Using this feature on the WMT 2010 French-English dataset, the authors report a gain of 0.55 BLEU points over a baseline phrase table. When added to modified Kneser-Ney, this feature produced an additional gain of .07 BLEU points.

4.1 METHODS

4.2 MAP

We extend MAP to an arbitrary number of corpora, M. In this formulation, models trained on corpora that are more distant from the test domain are successively MAP-adapted with models estimated from less distant corpora. This is done by sorting the corpora based on their distance to a development set and constructing an adaptation hierarchy as shown in Figure 1. The formulation of MAP for multiple corpora used in this paper is shown below:

$$\hat{p_i}(s|t,\lambda) = \frac{N_i(s,t)}{N_i(s,t)+\tau_i} p_i(s|t,\lambda_i) + \frac{\tau_i}{N_i(s,t)+\tau_i} \hat{p_{i+1}}(s|t,\lambda_{i+1}) \tag{6}$$

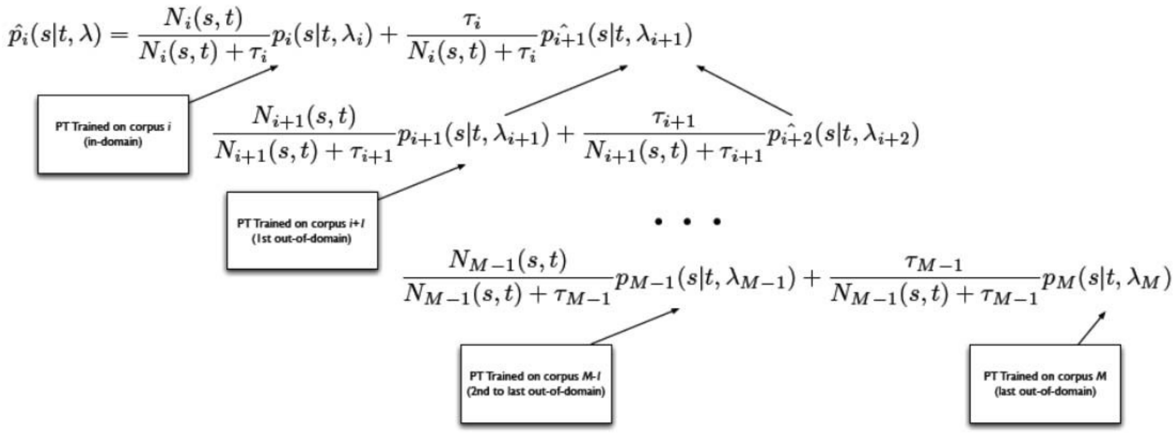

$\hat{p_i}(s|t,\lambda) = \frac{N_i(s,t)}{N_i(s,t)+\tau_i} p_i(s|t,\lambda_i) + \frac{\tau_i}{N_i(s,t)+\tau_i} \hat{p_{i+1}}(s|t,\lambda_{i+1})$
PT Trained on corpus *i* (in-domain)
$\frac{N_{i+1}(s,t)}{N_{i+1}(s,t)+\tau_{i+1}} p_{i+1}(s|t,\lambda_{i+1}) + \frac{\tau_{i+1}}{N_{i+1}(s,t)+\tau_{i+1}} \hat{p_{i+2}}(s|t,\lambda_{i+2})$
PT Trained on corpus *i+1* (1st out-of-domain)
• • •
$\frac{N_{M-1}(s,t)}{N_{M-1}(s,t)+\tau_{M-1}} p_{M-1}(s|t,\lambda_{M-1}) + \frac{\tau_{M-1}}{N_{M-1}(s,t)+\tau_{M-1}} p_M(s|t,\lambda_M)$
PT Trained on corpus *M-1* (2nd to last out-of-domain)
PT Trained on corpus *M* (last out-of-domain)

Figure 1: MAP with Multiple Corpora

MAP-based interpolation is done at the phrase-pair level, and the final probability estimate for the given phrase pair is $\hat{p}1\,(s|t)$.

We experiment with two methods of constructing the MAP hierarchy.

First, TFLLR-based similarity scores are computed using 1-gram document vectors created from the source language data from each corpus, with all words with *count* < 5 and all stop words removed. Background statistics are derived by concatenating all corpora used in these experiments. They are then used to compute a TFLLR weighting [12] which is applied to the document vectors. For each model, we compute the cosine similarity between the vectors from the training data for that model and from the development data. These similarities are then sorted to produce an ordering for MAP.

Second, we use each phrase table individually to translate the development dataset and create a MAP hierarchy from the resulting BLEU scores.

We present a method for determining τi based on the TFLLR-based similarity scores described above. τi is chosen so that the interpolation coefficient for each model is, on average, equal to the similarity score for that corpus, with the scores normalized to sum to one:

$$\alpha_i = s_i = \frac{\bar{N}_i}{\bar{N}_i + \tau_i} \qquad (7)$$

$$\tau_i = \frac{\bar{N}_i(1 - s_i)}{s_i} \qquad (8)$$

As described in 4.2, these normalized scores are also used as corpus weights for linear interpolation. With this choice of τi, the smoothing effect of MAP on the in-domain model is, on average, the same as the smoothing effect of linear interpolation with those weights.

4.3 Linear Interpolation

There are many ways to determine the coeffients used for linear interpolation, and no consensus on the best method of doing so [8]. For this paper, we compare the TFLLR-based similarity scores used to determine τ for MAP with uniform interpolation weights.

In both cases, linear interpolation has a smoothing effect similar to that produced by MAP estimation, but with a key difference. When a phrase exists only in the in-domain model, MAP reduces to:

$$\hat{p}(s|t,\lambda) = \frac{N_{adapt}(s,t)}{N_{adapt}(s,t)+\tau} p(s|t,\lambda_{adapt}) \tag{9}$$

MAP assumes that low-count phrase pairs are poorly estimated relative to higher-count pairs, and this smoothing effect removes more probability mass from the least reliable estimates. During linear interpolation, however, probabilities are trusted solely based on the domain used to estimate them, with no regard for the number of occurrences that produced that estimation.

Like the MAP smoothing effect, both Kneser-Ney smoothing and the ELF feature are count-based. Combining linear interpolation with these other smoothing methods may, therefore, result in additional performance improvements. We test this hypothesis by interpolating smoothed phrase tables. When combining linear interpolation with the ELF feature, we add one ELF feature to the phrase table for each corpus.

4.4 Phrase Table Fill-up

The rankings used to determine the MAP hierarchies are also used as the cascade orderings for phrase table fill-up. Fill-up interpolation has no smoothing effect, so we test it both with and without smoothing. When combining this interpolation method with the ELF feature, we use a slight modification of the original phrase table fill-up method. We use that ELF feature in place of the binary fill-up feature, and include an ELF feature for the in-domain model, to allow the optimizer to penalize low frequency words from that model. For comparison, we include a binary feature for the in-domain model in our baseline fill-up implementation.

5.1 Experiments

5.2 Data

All experiments performed for this paper, test English-French translation using four corpora: Europarl (EP), Gigaword, News Commentary (NC), and TED. The first three came from WMT 2011 [13], and the last from IWSLT 2011˜[14]. The NC and TED datasets are relatively small at just over 100K sentences each, while EPl contains 600K sentences and Gigaword 2.5M sentences.

We run all experiments on both the WMT 2011 and IWSLT 2011 test datasets. When testing on the WMT data, we use the WMT 2010 test set as development data for optimization. We consider the TED data to be the in-domain set for the IWSLT tests, and experiment with treating either the NC or EP data as in-domain for WMT.

5.3 System

Our baseline system uses a standard SMT architecture and has performed well on past evaluations [4,15].

We use interpolated Knesser-Ney n-gram language models built with the MIT Language Modeling Toolkit [16]. Additional class-based language models were also trained on the TED data and used for rescoring when translating the IWSLT dataset.

All phrase tables were created with IBM Model 4 alignments [17]; alignments extracted using the Berkeley Aligner and Competitive Linking Algorithm (CLA) were added to the NC and TED phrase tables [18].

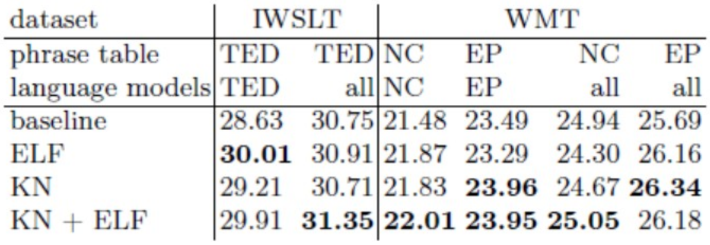
Our translation model assumes a log-linear combination of phrase translation models, language models, etc. We optimize the combination weights over a development set using minimum error rate training with a standard Powell-like grid search [19]. We use the Moses decoder [20].

All scores reported here are average BLEU scores obtained from three rounds of optimization.

5.4 Results

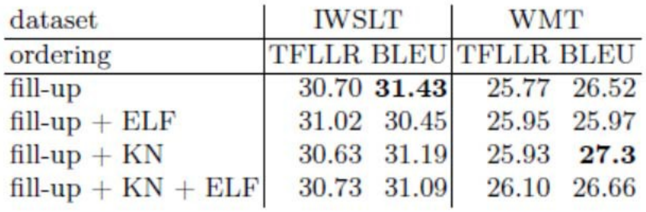
We ran four experiments exploring the effect of smoothing on a single phrase table: baseline, ELF, KN, and ELF + KN. We did two sets of these experiments: one in which only the in-domain phrase table and language model were used, and one in which the in-domain phrase table was paired with language models from all available corpora. Results are shown in Table 1. As expected, KN with the ELF feature has the best overall performance. The ELF feature seems to be particularly useful on the IWSLT dataset; the results are less consistent for WMT. These results suggest that the EP model should be considered in-domain for WMT. The extra language models provide a boost of 2 to 2.5 BLEU points across the board.

Table 1: Smothing Results



dataset	IWSLT		WMT			
phrase table	TED	TED	NC	EP	NC	EP
language models	TED	all	NC	EP	all	all
baseline	28.63	30.75	21.48	23.49	24.94	25.69
ELF	**30.01**	30.91	21.87	23.29	24.30	26.16
KN	29.21	30.71	21.83	**23.96**	24.67	**26.34**
KN + ELF	29.91	**31.35**	**22.01**	**23.95**	**25.05**	26.18

Table 2: Phrase Table Fill-Up Results



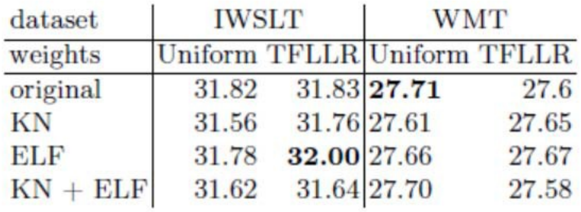
dataset	IWSLT		WMT	
ordering	TFLLR	BLEU	TFLLR	BLEU
fill-up	30.70	**31.43**	25.77	26.52
fill-up + ELF	31.02	30.45	25.95	25.97
fill-up + KN	30.63	31.19	25.93	**27.3**
fill-up + KN + ELF	30.73	31.09	26.10	26.66

We experimented with all of the above smoothing options in combination with both fill-up and distance-based interpolation. We used all four language models for all domain adaptation experiments.

Fill-up results are in Table 2. For the WMT dataset, fill-up provides a clear improvement over the baseline. Performance on the WMT dataset is strongly impacted by the fill-up cascade ordering, most likely because the TFLLR-based ordering treats the NC model as in-domain, while the BLEU ordering treats the EP model as in-domain.

Fill-up order has less of an impact on the IWSLT dataset; the TED model is considered in-domain for both. Regardless of order, fill-up performs badly on the IWSLT dataset, suggesting that some phrases from the out-of-domain corpora may be hurting performance.

Table 3: Linear Interpolation Results - IWSLT

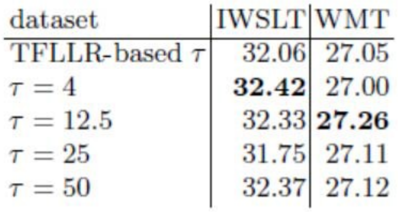


dataset	IWSLT		WMT	
weights	Uniform	TFLLR	Uniform	TFLLR
original	31.82	31.83	**27.71**	27.6
KN	31.56	31.76	27.61	27.65
ELF	31.78	**32.00**	27.66	27.67
KN + ELF	31.62	31.64	27.70	27.58

Linear interpolation results are in Table 3. For both datasets, the choice of weights had little impact on performance; smoothing also does not provide any additional improvement on top of interpolation. While linear interpolation is better than the baseline for both datasets, the gain is much larger for WMT.

MAP using the BLEU ordering outperformed MAP with the TFLLR-based ordering for both datasets. Using the BLEU ordering, we experimentedwith several constant values of MAP τ, as well as the TFLLR-based values described in Section 4.1. The results are shown in Table 4. Overall, MAP performance is relatively insensitive to the value of τ chosen.

Table 4: MAP Results



dataset	IWSLT	WMT
TFLLR-based τ	32.06	27.05
$\tau = 4$	**32.42**	27.00
$\tau = 12.5$	32.33	**27.26**
$\tau = 25$	31.75	27.11
$\tau = 50$	32.37	27.12

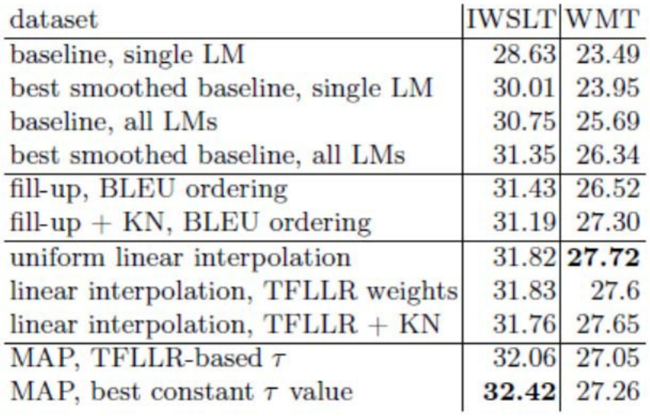
On the IWSLT dataset, MAP equals or outperforms linear interpolation for all values of τ, while linear interpolation consistently outperforms MAP on the WMT dataset.

The two datasets used here have very different characteristics. The TED training data is well-matched to the IWSLT test domain, but contains relatively few sentences. As a result, the model is poorly estimated, and benefits greatly from count-based smoothing. Data from other domains, however, does not consistently improve IWSLT performance. In this situation, MAP adaptation outperforms the other interpolation and smoothing techniques tested here.

The EP model, on the other hand, is trained on a large amount of data, but that data does not come from the WMT test domain. WMT performance is very much improved by the incorporation of out-of-domain data, and is less impactedby count-based smoothing techniques. Linear interpolation performs better than MAP adaptation in this case.

6.0 Conclusion

Table 5: Results Summary



dataset	IWSLT	WMT
baseline, single LM	28.63	23.49
best smoothed baseline, single LM	30.01	23.95
baseline, all LMs	30.75	25.69
best smoothed baseline, all LMs	31.35	26.34
fill-up, BLEU ordering	31.43	26.52
fill-up + KN, BLEU ordering	31.19	27.30
uniform linear interpolation	31.82	**27.72**
linear interpolation, TFLLR weights	31.83	27.6
linear interpolation, TFLLR + KN	31.76	27.65
MAP, TFLLR-based τ	32.06	27.05
MAP, best constant τ value	**32.42**	27.26

In this paper, we examined several methods of phrase table interpolation and compared them with the hierarchical MAP adaptation technique that we present. We also explore the relative contributions of smoothing and the addition of out-of-domain data to the performance gains achieved through phrase table interpolation.

For both the WMT 2011 and IWSLT 2011 datasets, phrase table fill-up performs worse than both linear interpolation and MAP adaptation. Linear interpolation is more effective than MAP on the WMT dataset, while the reverse is true for IWSLT.

These results can be explained by the nature of the datasets themselves. Count-based smoothing, and thus MAP adaptation, has a greater impact when the in-domain model is poorly estimated but well-matched to the test domain, as in the IWSLT dataset. When the converse is true, as in the WMT dataset, the incorporation of additional data provides the greatest performance improvement. Despite these differences, the MAP adaptation technique we present improves baseline performance by 1.5+ BLEU points on both datasets.

7.0 REFERENCES

1. Gauvain, J.L., Lee, C.H., Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing **2** (1994) 291–298
2. Federico, M., Bayesian Estimation Methods for N-Gram Language Model Adaptation. In Proceedings of International Conference on Spoken Language Processing. (1996) 240–243
3. Bacchiani, M., Riley, M., Roark, B., Sproat, R., Map Adaptation of Stochastic Grammars. Computer Speech & Language **20**(1) (2006) 41–68
4. Shen, W., Delaney, B., Aminzadeh, A.R., Anderson, T., Slyh, R., The MIT-LL/AFRL IWSLT-2009 System. Proc. of the International Workshop on Spoken Language Translation, Tokyo, Japan (2009) 71–78
5. Foster, G., Kuhn, R., Johnson, J.H.: Phrase Table Smoothing for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing, Sydney, Australia (2006)
6. Foster, G., Goutte, C., Kuhn, R., Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. Proceedings of the 2010 EMNLP, Cambridge, MA (October 2010) 451–459
7. Foster, G., Kuhn, R., Mixture-Model Adaptation for SMT. ACL Workshop on Statistical Machine Translation, Prague, Czech Republic (2007)
8. Bisazza, A., Ruiz, N., Federico, M., Fill-Up versus Interpolation Methods for Phrase-Based SMT Adaptation. International Workshop on Spoken Language Translation. (2011)
9. Kneser, R., Ney, H., Improved Backing-Off for M-Gram Language Modeling. Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference; **Volume 1**. (May 1995) 181 –184
10. Chen, S.F., Goodman, J., An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language **13**(4) (1999) 359 – 393
11. Chen, B., Kuhn, R., Foster, G., Johnson, H., Unpacking and Transforming Feature Functions: New Ways to Smooth Phrase Tables; Proceedings of MT Summit XIII, Xiamen, China (2011)
12. Campbell, W., Campbell, J., Gleason, T., Reynolds, D., Shen, W., Speaker Verification using Support Vector Machines and High-Level Features. Audio, Speech, and Language Processing, IEEE Transactions on **15**(7) (sept. 2007) 2085 –2094
13. Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O., Findings of the 2011 Workshop on Statistical Machine Translation. Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland (July 2011) 22–64
14. Federico, M., Luisa Bentivogli Michael Paul, S.S.: Overview of the IWSLT 2011 Evaluation Campaign. Proceedings of the International Workshop on Spoken Language Translation, San Francisco, CA (November 2011)
15. Shen, W., Anderson, T., Slyh, R., Aminzadeh, A., The MIT-LL/AFRL IWSLT-2010 MT System. Proc. of the International Workshop on Spoken Language Translation, Paris, France

(2010)

16. Hsu, B.J., Glass, J., Iterative Language Model Estimation: Efficient Data Structure and Algorithms. In Proc. Interspeech. (2008)
17. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L., The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics **19**(2) (1993) 263–311
18. Chen, B., Cattoni, R., Bertoldi, N., Cettolo, M., Federico, M., The ITC-IRST SMT System for IWSLT-2005. Proceedings of the IWSLT 2005. (2005)
19. Och, F.J., Minimum Error Rate Training in Statistical Machine Translation. Proceedings of ACL. (2003)
20. Koehn, P., Hoang, H., Birch, A., Burch, C.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. ACL ’07, Stroudsburg, PA, USA (2007) 177–180