MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation

Nick Ruiz, Marcello Federico

FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38123 Povo (TN), Italy
{nicruiz,federico}@fbk.eu

Abstract

This paper provides a fast alternative to Minimum Discrimination Information-based language model adaptation for statistical machine translation. We provide an alternative to computing a normalization term that requires computing full model probabilities (including back-off probabilities) for all n-grams. Rather than re-estimating an entire language model, our Lazy MDI approach leverages a smoothed unigram ratio between an adaptation text and the background language model to scale only the n-gram probabilities corresponding to translation options gathered by the SMT decoder. The effects of the unigram ratio are scaled by adding an additional feature weight to the log-linear discriminative model. We present results on the IWSLT 2012 TED talk translation task and show that Lazy MDI provides comparable language model adaptation performance to classic MDI.

1. Introduction

Topic adaptation is used as a technique to adapt language models based on small contexts of information that may not necessarily reflect an entire domain or genre. In scenarios such as lecture translation, it is advantageous to perform language model adaptation on the fly to reflect topical changes in a discourse. In these scenarios, general purpose domain adaptation techniques fail to capture the nuances of discourse; while domain adaptation works well in modeling newspapers and government texts which contain a limited number of subtopics, the genres of lectures and speech may cover a virtually unbounded number of topics that change over time. Instead of general purpose adaptation, adaptation should be performed on smaller windows of context.

Most domain adaptation techniques require the re-estimation of an entire language model to leverage the use of out-of-domain corpora in the construction of robust models. While efficient algorithms exist for domain adaptation, they are in practice intended to adapt language models globally over a new translation task. Topic adaptation, on the other hand, intends to adapt language models as relevant contextual information becomes available. For a speech, the relevant contextual information may come in sub-minute intervals. Well-established and efficient techniques such as Minimum Discrimination Information adaptation [1, 2] are unable to perform topic adaptation in real-time scenarios for large order n-gram language models. In practice, new contextual information is likely to be available before techniques such as MDI have finished LM adaptation from earlier contexts. Thus spoken language translation systems are typically unable to use the state-of-the-art techniques for the purpose of topic adaptation.

In this paper, we seek to apply MDI adaptation techniques in real-time translation scenarios by avoiding the computation of the normalization term that requires all n-grams to be re-estimated. Instead, we only wish to adapt n-grams that appear within an adaptation context. Dubbed "Lazy MDI", our technique uses the same unigram ratios as MDI, but avoids normalization by applying smoothing transformations based a sigmoid function that is added as a new feature to the conventional log-linear model of phrase-based statistical machine translation (SMT). We observe that Lazy MDI performs comparably to classic MDI in topic adaptation for SMT, but possesses the desired scalability features for real-time adaptation of large-order n-gram LMs.

This paper is organized as follows: In Section 2, we discuss relevant previous work. In Section 3, we review MDI adaptation. In Section 4, we describe Lazy MDI adaptation for machine translation and review how unigram statistics of adaptation texts can be derived using bilingual topic modeling. In Section 5, we report adaptation experiments on TED talks[1] from IWSLT 2010 and 2012, followed by our conclusions and suggestions for future work in Section 6.

2. Previous Work

This paper is based on the work of [3], which combines MDI adaptation with bilingual topic modeling on small adaptation contexts for lecture translation. Adaptation texts are drawn from source language input and leveraged for language model adaptation. A bilingual Probabilistic Latent Semantic Analysis (PLSA) [4] model is constructed by combining parallel training texts, allowing for inference on monolingual source texts for MDI adaptation by removing source language unigram statistics.

[1] http://www.ted.com/talks

A similar approach is considered by [5] in domain adaptation by constructing two hierarchical LDA models from parallel document corpora and enforcing a one-to-one correspondence between the models by learning the hyperparameters of the variational Dirichlet posteriors in one LDA model and bootstrapping the second model by fixing the hyperparameters. The bilingual LSA framework is also applied to adapt translation models. Other bilingual topic modeling approaches include Hidden Markov Bilingual Topic AdMixtures [6] and Polylingual Topic Models [7].

The literature focuses primarily on domain adaptation, using techniques such as information retrieval to select similar sentences in training corpora for adaptation, either through interpolation [8] or corpora filtering [9], or mixture model adaptation approaches [10, 11].

An alternative to MDI adaptation is proposed by [12], which uses a log-linear combination of binary features $f_i(h, w)$ to scale LM probabilities $P(w \mid h)$:

$$\hat{P}(w \mid h) = \exp\left(\sum_i f_i(h, w)\lambda_i\right) P(w \mid h).$$

Normalization is avoided by simply dividing $\hat{P}(w \mid h)$ by $\hat{P}(w \mid h) + 1$.

3. MDI Adaptation

MDI adaptation was originally presented in [1] as a means for domain adaptation on language models. MDI adaptation scales the probabilities of a background language model, $P_B(h, w)$, by a factor determined by a ratio between the unigram statistics observed in an adaptation text A versus the same statistics observed in the background corpus B:

$$\alpha(w) = \left(\frac{\hat{P}_A(w)}{P_B(w)}\right)^\gamma, \quad 0 < \gamma \leq 1. \tag{1}$$

As such, the adapted language model $P_A(h, w)$ is constructed as follows:

$$P_A(h, w) = P_B(h, w)\alpha(w), \tag{2}$$

where h is the n-gram history of word w. As outlined in [13], the adapted language model can also be written recursively in an interpolated conditional form with discounted frequencies $f^*(w|h)$ and reserved probabilities for out-of-vocabulary words $\lambda(h)$:

$$P_A(w|h) = f_A^*(w|h) + \lambda_A(h)P_A(w|h'), \tag{3}$$

with:

$$f_A^*(w|h) = \frac{f_B^*(w|h)\alpha(w)}{z(h)}, \tag{4}$$

$$\lambda_A(h) = \frac{\lambda_B(h)z(h')}{z(h)}, \tag{5}$$

and

$$z(h) = \Big(\sum_{w:N_B(h,w)>0} f_B^*(w|h)\alpha(w)\Big) + \lambda_B(h)z(h'), \tag{6}$$

which efficiently computes the normalization term for high order n-grams recursively by just summing over observed n-grams. The recursion ends with the following initial values for the empty history ϵ:

$$z(\epsilon) = \sum_w P_B(w)\alpha(w), \tag{7}$$

$$P_A(w|\epsilon) = P_B(w)\alpha(w)z(\epsilon)^{-1}. \tag{8}$$

While MDI has been applied in domain adaptation both for language models [2] and translation models [5], its re-estimation requires the computation of the normalization term outlined in (6). In topic adaptation scenarios, it is desirable to rapidly adapt a background language model using small adaptation contexts consisting of few sentences. One method of inferring unigram statistics for MDI adaptation given sparse data is to perform bilingual topic modeling [3, 5, 7]. While it has been shown that the combination of topic modeling and MDI adaptation yield a significant improvement in translation adequacy, the approach of adapting non-overlapping contexts of size C requires M/C full LM re-estimations on a translation task with M sentences, with each re-estimation requiring the expensive computation of the normalization term.

4. Lazy MDI Alternative for SMT

The goal of MDI adaptation is to construct an adapted language model that minimizes its Kullback-Leibler divergence from the background LM, which is effectively performed via the unigram ratio scaling method described in (1) and (2). We seek to loosely approximate this KL divergence in statistical machine translation by adapting only n-grams that appear as translation options for a given sentence. As such, we seek to avoid computing a normalization term that requires observing the probabilities of all high- and lower-order n-grams in the LM. Since the ratio of unigram probabilities is defined across the range $[0, +\infty]$, we explore smoothing functions that bind the ratio to a finite range.

4.1. Smoothing unigram ratios

In machine learning, sigmoid activation functions are typically used to constrain functions in the range of $[0, a]$ or $[-a, a]$ to reduce the bias of a few data points within a training set. Likewise we explore the use of sigmoid functions to reward n-gram probabilities across the range of $[0, a]$. However, since we are scaling ratios in general, we desire the following properties of our smoothing function f:

$$f(0) = 0; \quad \lim_{x\to+\infty} f(x) = a$$

$$f(1) = 1; \quad \lim_{x\to-\infty} f(x) = -a$$

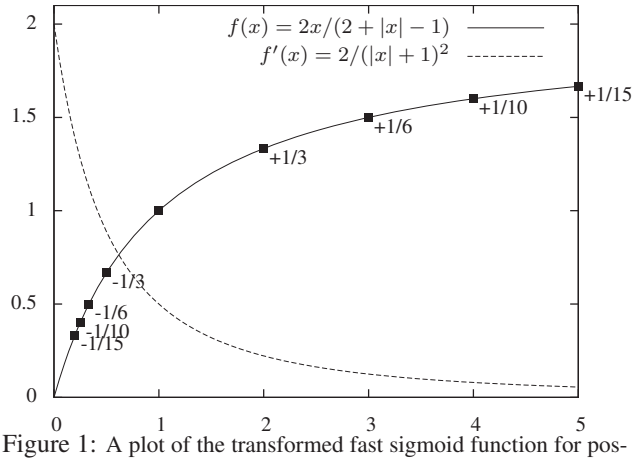

Figure 1: A plot of the transformed fast sigmoid function for positive ratios in (10) and its first derivative, evaluated at $a = 2$. The relative changes in $f(x)$ are labeled, centered at $f(1)$. The changes in $f(x)$ are symmetric with respect to each ratio and inverse ratio.

In particular the $f(1) = 1$ constraint ensures that background LM probabilities remain fixed when the ratio is balanced.

Staple sigmoid functions such as the logistic function or the hyperbolic tangent unfortunately cannot satisfy the property $f(1) = 1$ for any magnitude a. However, a *fast sigmoid* approximation was proposed in [14], defined as:

$$f(x) = \frac{x}{1 + |x|}. \tag{9}$$

With some simple transformations, we arrive at our desired function:

$$f(x, a) = \frac{ax}{a + |x| - 1}, \quad a > 1. \tag{10}$$

Figure 1 contains a plot of (10) at $a = 2$ and its first derivative. A useful property of the fast sigmoid in (10) is that the change in slope is symmetric with respect to inverted ratios, relative to the center at $x = 1$. For example, for the fast sigmoid outlined in Figure 1, a ratio of 2:1 yields a scale of $1 + \frac{1}{3}$, while a ratio of 1:2 yields a scale of $1 - \frac{1}{3}$.

4.2. Log-linear feature

Since we are no longer normalizing n-gram probabilities, we can consider the smoothed unigram probabilities as a function that rewards or penalizes translation options based on the likelihood that the words composing the target phrase should appear in the translation. We treat the smoothed unigram probabilities as a new feature in the discriminative log-linear model of the decoder. While our new feature is independent from any language model features, we can logically consider the adaptation of a background language model as a log-linear combination of the LM feature and the Lazy MDI feature as:

$$\hat{P}_{LM}(E \mid F) = P_{LM}(E \mid F)^{\gamma_1} \cdot \prod_{i=1}^{|E|} \hat{\alpha}(e_i)^{\gamma_2}, \tag{11}$$

where $P_{LM}(E \mid F)$ computes the language model probabilities of target sentence E, given a source sentence F; though we only consider language models that score the target sentence, independent from F. $\hat{\alpha}(e_i)$ is the Lazy MDI adaptation on the ith target word in E, defined as:

$$\hat{\alpha}(w) = f\left(\frac{P_A(w)}{P_B(w)}\right). \tag{12}$$

By rearranging terms, we arrive at our unnormalized log-linear approximation of (2):

$$\hat{P}_{LM}(E) = \prod_{i=1}^{|E|} P_{LM}(e_i \mid h_i)^{\gamma_1} \cdot \hat{\alpha}(e_i)^{\gamma_2}. \tag{13}$$

In practice, only translation hypotheses suggested by the translation model are scored by the language model, thus limiting the number of unigram ratios to consider. Additionally, for computational efficiency, calculations are performed in log space. For $a = 2$, our fast sigmoid function can be rewritten as:

$$f(x, 2) = 2 \cdot \left(1 + e^{-\ln(x)}\right)^{-1}, \quad x > 0, \tag{14}$$

which allows us to compute log probability ratios as $\ln P_A(w) - \ln P_B(w)$.

4.3. Sparsity considerations

If we treat the background and adaptation unigram statistics as unigram language models, we can use smoothing to reserve probability for out-of-vocabulary words. However, due to the sparsity of unigram features in adaptation texts, it is possible that the adapted unigram statistics are missing words that appear in the background LM. Assuming that there are insufficient adaptation statistics to reliably scale the probabilities of n-grams containing these words, we instead leave the background probabilities intact by fixing the unigram probability ratio to 1.

A similar problem can arise in the scenario that the adaptation text contains unigrams that are not observed in the background LM. One possible solution is to limit the vocabulary of the adaptation statistics to the same as that of the background.

4.4. Inferring unigrams via bilingual topic modeling

Since an adaptation text is in practice too small to directly compute reliable unigram statistics, we resort to topic modeling approaches to infer full unigram probabilities. One such approach is Probabilistic Latent Semantic Analysis (PLSA) [4], which computes the probability of unigrams in a document d by marginalizing over a collection of latent topics Z:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d). \tag{15}$$

Following the exposition of [3], we construct a bilingual topic model by combining source and target parallel

sentences into "monolingual" documents with vocabulary $V_{FE} = V_F \cup V_E$.[2] During inference, we infer unigram probabilities of V_{FE} using only documents containing only the source language. Removing words $f \in V_F$ from the probability distribution and normalizing yields a probability distribution for all words in V_E.

5. Experiments

We conduct experiments on the IWSLT TED talk translation tasks from 2010 and 2012. In Section 5.1, we evaluate the utility of Lazy MDI using lowercased unigram statistics on a lowercased MT system trained only on TED data. We compare the performance of smoothed and unsmoothed Lazy MDI against classic MDI.

In Section 5.2, we evaluate the logical adaptation of cased language models with uncased unigram statistics from both the adaptation text and the background text. Due to the small size of the adaptation texts, we are not guaranteed a reliable unigram probability estimations on a vocabulary that is likely to double in size. We evaluate the utility of Lazy MDI on a state-of-the-art system against a domain-adapted mixture LM.

5.1. IWSLT 2010

We replicate the experimental settings of [3] and provide a comparison of classic MDI against Lazy MDI, using the same data set of English-French translations of TED talks, downloaded from the TED website as it was on March 30, 2011 and split into training, dev and test sets according to indexes used for IWSLT 2010[3] evaluation. The data set is segmented at the clause level, rather than at the level of sentences. The TED training data consists of 329 parallel talk transcripts with approximately 84k sentences. The TED test data consists of transcriptions created via 1-best ASR outputs from the KIT Quaero Evaluation System. It consists of 2381 clauses and approximately 25,000 English and French words, respectively.

Lowercased SMT systems are built upon the Moses open-source SMT toolkit [15][4]. The translation and lexicalized reordering models have been trained on parallel data. One 5-gram background LM was constructed with the IRSTLM toolkit [16] on the French side of the TED training data (740k words), and smoothed via the improved Kneser-Ney technique [17]. The weights of the log-linear interpolation model were optimized via minimum error rate training (MERT) [18] on the TED development set, using 200 best translations at each tuning iteration.

As in [3], online adaptation is simulated by splitting the training corpus into small non-overlapping contexts of 5 lines (41,847 "documents" in total) and performing bilingual PLSA training using IRSTLM. The PLSA model consists of 250 topics and is trained for 20 EM iterations. Ten inference iterations are performed on the English side of the development and test sets to generate French unigram probabilities for each 5-line context.

MDI adaptation is performed on the test set contexts using the 5-gram TED language model described above as the background. For each 5-line context in the test set, the background LM is replaced with the adapted LM for SMT decoding, preserving the same feature weight as the background LM.

In the case of Lazy MDI, adaptation is integrated into the Moses decoder using the same context unigrams. MERT is performed on the development set with simultaneous adaptation for each context. We experiment with both adaptation via unsmoothed unigram ratios and smoothing via our transformed fast sigmoid function. Words not in the adaptation unigram LM are fixed with a 1:1 ratio to prevent their effect on the global translation hypothesis score.

We ran 3 MERT instances for each system and evaluated using MultiEval 0.3 [19]. Evaluation results in terms of BLEU, METEOR (French), TER, and segment length are listed in Table 1. We observe similar results between MDI and smoothed Lazy MDI – both of which yield an average improvement of 0.2 and 0.3 BLEU, respectively. As predicted, unsmoothed Lazy MDI adaptation performs poorly as the unigram ratios between the background and context LMs often diverge greatly. This can also be observed in the weight associated with the feature, as shown in Table 2. For unsmoothed Lazy MDI, the associated feature weight has divergent values across each MERT instance, implying the un-

Metric	System	Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p
BLEU ↑	Baseline	28.0	0.5	0.3	-
	MDI	28.2	0.5	0.2	0.01
	Lazy MDI (unsmoothed)	24.4	0.5	5.8	0.00
	Lazy MDI (smoothed)	28.3	0.5	0.1	0.00
METEOR ↑	Baseline	50.4	0.4	0.1	-
	MDI	50.6	0.5	0.2	0.09
	Lazy MDI (unsmoothed)	47.7	0.4	4.3	0.00
	Lazy MDI (smoothed)	50.5	0.4	0.1	0.18
TER ↓	Baseline	57.3	0.6	0.4	-
	MDI	56.9	0.6	0.4	0.00
	Lazy MDI (unsmoothed)	61.9	0.6	8.0	0.00
	Lazy MDI (smoothed)	56.9	0.6	0.1	0.00
Length	Baseline	104.1	0.5	1.1	-
	MDI	103.5	0.5	0.9	0.00
	Lazy MDI (unsmoothed)	106.2	0.5	4.5	0.00
	Lazy MDI (smoothed)	103.5	0.5	0.2	0.00

Table 1: Lowercased evaluation of MDI and Lazy MDI adaptation techniques on the IWSLT 2010 TED test set. Metric scores averaged across three MERT runs. p-values are relative to the baseline. s_{sel} indicates the variance due to test set selection. Significant improvements in terms of BLEU and TER are observed for both MDI and smoothed Lazy MDI (via a fast sigmoid transformation of unigram ratios). Unsmoothed Lazy MDI yields unpredictable results during optimization.

[2] To avoid overlapping types in the topic mdoel, we annotate the source and target vocabularies to track their provenance.

[3] http://iwslt2010.fbk.eu/

[4] http://www.statmt.org/moses/

predictability of unbounded ratios.

System	Metric	Opt 1	Opt 2	Opt 3
Baseline	BLEU	27.64	28.20	28.20
MDI	BLEU	28.49	28.07	28.16
Lazy MDI (unsmoothed)	BLEU	27.14	*17.80*	28.40
	weight	0.1537	*0.4096*	0.0445
Lazy MDI (smoothed)	BLEU	28.27	28.39	28.17
	weight	0.0132	0.0177	0.0138

Table 2: Lowercased evaluation runs for the TED baseline and Lazy MDI adaptations for the IWSLT 2010 test set across three tuning instances. Unsmoothed Lazy MDI yields unstable adaptation feature weights across each run. "Opt 2" overpowers the log-linear model, causing a large overfitting to the development set. "Opt 3" provides the best generalization to the test set by reducing the effects of the adaptation. For fast sigmoid-smoothed Lazy MDI, the adaptation weights remain consistent across all runs.

5.2. IWSLT 2012

We also evaluate the performance of our fast sigmoid-smoothed Lazy MDI setting on a state-of-the-art SMT system submitted for the IWSLT 2012 TED English-French MT shared task[5]. In this experiment, we build cased translation systems using Moses and evaluate the effects of Lazy MDI adaptation from lowercased unigram context statistics. Our baseline system consists of translation and reordering models trained from the in-domain TED[6] corpus, as well as out-of-domain Giga French-English[7] and Europarl v7 [21] corpora. Each out-of-domain corpus was domain-adapted by aggressive filtering using a cross-entropy difference scoring technique described by [22] on the French side and optimizing the perplexity against the (French) TED training data by incrementally adding sentences. The corresponding parallel English sentences were preserved to provide compact parallel corpora. A single phrase and reordering table were constructed using the fill-up technique described in [23] in a cascaded fashion in the order of TED, Giga French-English, and Europarl.

A domain-adapted 5-gram mixture language model was constructed with IRSTLM from the TED, Giga French-English, Gigaword French v2 AFP[8], and WMT News Commentary v7 corpora. The same filtering technique [22] was applied to the LM corpora. For Lazy MDI, we again use the bilingual PLSA model constructed from the IWSLT 2010 training data, with 250 topics and 20 EM iterations. MERT is again performed on the development set with simultaneous Lazy MDI adaptation for each context.

Topic adaptation results against the domain-adapted baseline are shown in Table 3. The evaluation results are averaged over three MERT optimizations of the baseline and Lazy MDI-adapted systems. We observe that performing Lazy MDI adaptation yields a BLEU improvement of 0.2 against the already-adapted baseline, suggesting a cumulative gain of domain adaptation and topic adaptation. We also observe a 0.2 improvement in terms of TER, while METEOR remains more or less the same. The tuning weights obtained across three MERT iterations are averaged to control optimizer instability. We list the evaluation results of each system run in Table 4.

Metric	System	Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p
BLEU ↑	Mix LM	32.4	0.5	0.0	-
	+Lazy MDI	32.6	0.5	0.1	0.07
METEOR ↑	Mix LM	52.0	0.4	0.0	-
	+Lazy MDI	52.1	0.4	0.1	0.18
TER ↓	Mix LM	49.5	0.5	0.1	-
	+Lazy MDI	49.3	0.5	0.2	0.05
Length	Mix LM	97.3	0.4	0.3	-
	+Lazy MDI	97.2	0.4	0.2	0.12

Table 3: Evaluation of Lazy MDI adaptation on the IWSLT 2010 TED test set provided in the IWSLT 2012 TED translation task. Metric scores averaged across three MERT runs. Lazy MDI p-values are relative to the domain-adapted baseline, described in Section 5.2. s_{sel} indicates the variance due to test set selection. Significant improvements in terms of BLEU and TER are observed for smoothed Lazy MDI (via a fast sigmoid transformation of unigram ratios).

System	Metric	Opt 1	Opt 2	Opt 3	Avg
Mix LM	BLEU	32.37	32.44	32.44	32.42
	NIST	7.463	7.438	7.438	7.443
+Lazy MDI	BLEU	32.63	32.55	32.52	32.70
	NIST	7.473	7.480	7.440	7.448

Table 4: Lowercased evaluation runs for the mixture LM baseline and Lazy MDI adaptations for the 2010 test set in the IWSLT 2012 translation task, across three tuning instances. The weights from the tuning instances are averaged to control optimizer instability. Performing Lazy MDI adaptation on the mixture LM baseline yields a 0.28 BLEU improvement and marginal NIST improvements.

We evaluate the impact of Lazy MDI adaptation by computing TER on the translation of each individual line from the 2010 test set by each system. We observe that of the 1,664 transcript lines, 247 lines yield a TER improvement, while 175 result in a higher error rate. We show three examples of segments yielding a TER improvement in Table 5. For ID #364, Lazy MDI yields a slight increase in fluency, while adequacy remains more or less the same. The baseline suggests that white pills are worse than blue pills – a subtle difference from the intent of the reference. The Lazy-adapted hypothesis corrects this difference, but makes common mistakes in translating "good" and "as". Lazy MDI yields a shorter translation in ID #1055 that moves away from a literal translation in the first half of the sentence that closely matches the reference. ID #1059 results in a very minor article change from

[5]http://hltc.cs.ust.hk/iwslt/index.php/evaluation-campaign/ted-task

[6]https://wit3.fbk.eu/mt.php?release=2012-03-test

[7]10^9 French-English data set provided by the WMT 2012 translation task [20].

[8]http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T28

"the" to "our". In this context, this subtle difference is important because the speaker is comparing the water at his fish farm to other farms.

ID	Text	TER
364	But a white pill is not as good as a blue pill .	
	Mais un comprimé blanc n' est pas aussi bon qu' une comprimé bleu	(0.154)
	Mais une pilule blanche est moins bonne qu' une pilule bleue .	0.769
	Mais une pilule blanche n' est pas aussi bien comme une pilule bleue .	0.615
1055	I mentioned that to Miguel , and he nodded .	
	J' ai dit ça à Miguel , et il a acquiesçé .	(0.167)
	J' ai mentionné que de Miguel , et il a fait un signe .	0.500
	J' ai dit à Miguel , et il a fait un signe .	0.333
1059	And then he added , " But our water has no impurities . "	
	Et puis il a ajouté : " Mais notre eau n' a pas d' impuretés . "	(0.058)
	Et puis il a ajouté : " Mais l' eau n' a pas impuretés . "	0.176
	Et puis il a ajouté : " Mais notre eau n' a pas impuretés . "	0.118

Table 5: Three examples of improvement in MT results: the first translation in each collection corresponds to the reference translation, the second utilizes a mixture LM, and the third adds Lazy MDI adaptation. The sentence-level TER scores are listed by each hypothesis and the difference is listed in parentheses by the reference.

We also outline three examples of diminished performance after performing Lazy MDI in Table 6. The Lazy MDI example in ID #858 demonstrates an attempt to literally translate the word "space" as "espace", which can ambiguously refer either to outer space, or a domain (as in the reference translation). This surface word is likely to have been chosen above "domaine" due to its topic similarity to "nucléaire". While the TER on this sentence is higher than the baseline, it should be noted that the baseline didn't provide a translation for "space". ID #895 is an example where the topic adaptation attempts to literally translate "I think", but adds an additional "that" afterward. The sentence becomes a bit awkward to read. The baseline, however, leaves out the hedge phrase "I think" and comes across as factual. It is likely that a human translator would prefer the topic-adapted sentence. In ID #1358, synonyms for "globe" are selected, correctly implying that the speaker refers to a globe as the world. While the reference and baseline select the word "planet", the topic-adapted sentence prefers "world" – an equally acceptable word. It is likely that "world" was selected due to collocations with "trash" and "pollution". With only one reference translation, it is hard to detect when Lazy MDI adaptation actually worsens the translation hypothesis.

6. Conclusions

We have presented a simplified framework for approximating MDI adaptation in an online manner for lecture translation. We avoid normalization computations that prevent

ID	Text	TER
858	In the nuclear space , there are other innovators .	
	Dans le domaine nucléaire , il y a d' autres innovateurs .	(-0.167)
	Dans le nucléaire , il y a d' autres innovateurs .	0.083
	En l' espace nucléaire , il y a d' autres innovateurs .	0.250
895	And so there is a thread of something that I think is appropriate .	
	Mais là-dedans , il y a quelque chose qui ne me semble pas faux .	(-0.267)
	Et il y a un fil de quelque chose qui est approprié .	0.600
	Et il y a un fil de quelque chose que je pense que c' est approprié .	0.867
1358	and not only that , we 've used our imagination to thoroughly trash this globe .	
	Pire , nous avons utilisé notre imagination pour polluer profondément cette planète .	(-0.154)
	Et non seulement ça , nous avons utilisé notre imagination à ordures soigneusement cette planète .	0.538
	Et non seulement ça , nous avons utilisé notre imagination à ordures soigneusement ce monde .	0.692

Table 6: Three examples of decreased TER performance in MT results: the first translation in each collection corresponds to the reference translation, the second utilizes a mixture LM, and the third adds Lazy MDI adaptation. The sentence-level TER scores are listed by each hypothesis and the difference is listed in parentheses by the reference.

classic MDI from being used in speech translation scenarios. Lazy MDI adaptation acts as a separate log-linear feature that doesn't directly adapt LM probabilities – instead, it rewards or penalizes the scores of each translation hypothesis by observing the unigram probabilities inferred an adaptation context and compares it to the background in a smoothed ratio. The smoothing is performed by a conservative fast sigmoid function that favors 1:1 ratios and prevents ratios from growing above a magnitude a.

We conducted adaptation experiments on TED talk data from IWSLT 2010 and 2012 and demonstrate a significant improvement in terms of BLEU, NIST, and TER over two baselines: a lowecased TED-only system, and a state-of-the-art cased system that combines in-domain and out-of-domain data. We demonstrate that Lazy MDI adaptation has cumulative adaptation effects on already-adapted language models.

For future work, we intend to compare our fast sigmoid function against non-sigmoidal smoothing functions for Lazy MDI. We additionally intend to explore log-linear alternatives that do not rely on the computation of unigram ratios – for example, inferring context from semantically-rich resources, such as Wikipedia or WordNet.

As it currently stands, Lazy MDI adaptation scales unigram ratios from data sources with differing vocabularies. It is likely that we can gain more reliable ratios by filtering the background unigram LM vocabulary to match the adaptation text and renormalizing the probabilities.

Another potential weakness in our approach is the use of topic models that do not filter stop-words and perform unigram adaptation on the surface level. For morphologically-

rich languages, such as German or Arabic, the vocabulary sizes can increase greatly due to word splitting. We intend to test our adaptation approach using word stems.

7. Acknowledgements

This work was supported by the T4ME network of excellence (IST-249119), funded by the DG INFSO of the European Commission through the Seventh Framework Programme.

8. References

[1] S. A. Della Pietra, V. J. Della Pietra, R. Mercer, and S. Roukos, "Adaptive language model estimation using minimum discrimination estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, San Francisco, CA, 1992, pp. 633–636.

[2] M. Federico, "Efficient language model adaptation through MDI estimation," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 4, Budapest, Hungary, 1999, pp. 1583–1586.

[3] N. Ruiz and M. Federico, "Topic adaptation for lecture translation through bilingual latent semantic models," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 294–302. [Online]. Available: http://www.aclweb.org/anthology/W11-2133

[4] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proceedings of the 15th Conference on Uncertainty in AI*, Stockholm, Sweden, 1999, pp. 289–296.

[5] Y.-C. Tam, I. Lane, and T. Schultz, "Bilingual LSA-based adaptation for statistical machine translation," *Machine Translation*, vol. 21, pp. 187–207, December 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1466799.1466803

[6] B. Zhao and E. P. Xing, "HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 1689–1696.

[7] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual Topic Models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, August 2009.

[8] B. Zhao, M. Eck, and S. Vogel, "Language Model Adaptation for Statistical Machine Translation via Structured Query Models," in *Proceedings of Coling 2004*. Geneva, Switzerland: COLING, Aug 23–Aug 27 2004, pp. 411–417.

[9] A. Sethy, P. Georgiou, and S. Narayanan, "Selecting relevant text subsets from web-data for building topic specific language models," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 145–148. [Online]. Available: http://www.aclweb.org/anthology/N/N06/N06-2037

[10] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0217

[11] P. Koehn and J. Schroeder, "Experiments in Domain Adaptation for Statistical Machine Translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 224–227. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0233

[12] S. F. Chen, K. Seymore, and R. Rosenfeld, "Topic adaptation for language modeling using unnormalized exponential models," in *IEEE ICASSP-98*. IEEE, 1998, pp. 681–684.

[13] M. Federico, "Language Model Adaptation through Topic Decomposition and MDI Estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, Orlando, FL, 2002, pp. 703–706.

[14] G. M. Georgiou, "Parallel distributed processing in the complex domain," Ph.D. dissertation, Tulane University, New Orleans, LA, USA, 1992, uMI Order No. GAX92-29796.

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

[16] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.

[17] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.

[18] F. J. Och, “Minimum Error Rate Training in Statistical Machine Translation,” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021.pdf

[19] J. Clark, C. Dyer, A. Lavie, and N. Smith, “Better hypothesis testing for statistical machine translation: Controlling for optimizer instability,” in *Proceedings of the Association for Computational Lingustics*, ser. ACL 2011. Portland, Oregon, USA: Association for Computational Linguistics, 2011, available at http://www.cs.cmu.edu/ jhclark/pubs/significance.pdf.

[20] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, “Findings of the 2012 workshop on statistical machine translation,” in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 10–51. [Online]. Available: http://www.aclweb.org/anthology/W12-3102

[21] P. Koehn, “Europarl: A multilingual corpus for evaluation of machine translation,” Unpublished, `http://www.isi.edu/~koehn/europarl/`, 2002.

[22] R. C. Moore and W. Lewis, “Intelligent selection of language model training data,” in *ACL (Short Papers)*, 2010, pp. 220–224.

[23] A. Bisazza, N. Ruiz, and M. Federico, “Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation,” in *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.