Monolingual Marginal Matching for Translation Model Adaptation

Ann Irvine
Johns Hopkins University
anni@jhu.edu

Chris Quirk
Microsoft Research
chrisq@microsoft.com

Hal Daumé III
University of Maryland
me@hal3.name

Abstract

When using a machine translation (MT) model trained on OLD-domain parallel data to translate NEW-domain text, one major challenge is the large number of out-of-vocabulary (OOV) and new-translation-sense words. We present a method to identify new translations of both known and unknown source language words that uses NEW-domain comparable document pairs. Starting with a joint distribution of source-target word pairs derived from the OLD-domain parallel corpus, our method recovers a new joint distribution that matches the marginal distributions of the NEW-domain comparable document pairs, while minimizing the divergence from the OLD-domain distribution. Adding learned translations to our French-English MT model results in gains of about 2 BLEU points over strong baselines.

1 Introduction

When a statistical machine translation (SMT) model trained on OLD-domain (e.g. parliamentary proceedings) parallel text is used to translate text in a NEW-domain (e.g. medical or scientific), performance degrades drastically. One of the major causes is the large number of NEW-domain words that are out-of-vocabulary (OOV) with respect to the OLD-domain text. Figure 1 shows the OOV rate for text in several NEW-domains, with respect to OLD-domain parliamentary proceedings. Even more challenging are the difficult-to-detect new-translation-sense (NTS) words: French words that are present in both the OLD and NEW domains but that are translated differently in each domain. For example, the French word *enceinte* is mostly translated in parliamentary proceedings as *place*, *house*, or *chamber*; in medical text, the translation is mostly *pregnant*; in scientific text, *enclosures*.

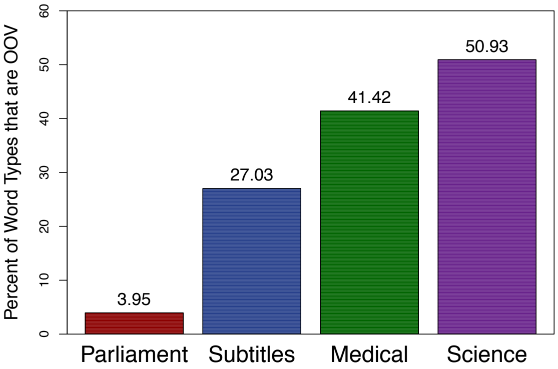


Figure 1: Percent of test set word types by domain that are OOV with respect to five million tokens of OLD-domain French parliamentary proceedings data.

One potential remedy is to collect parallel data in the NEW-domain, from which we can train a new SMT model. Smith et al. (2010), for example, mine parallel text from comparable corpora. Parallel sentences are informative but also rare: in the data released by Smith et al. (2010), only 21% of the foreign sentences have a near-parallel counterpart in the English article.[1] Furthermore, these sentences do not capture all terms. In that same dataset, we find that on average only 20% of foreign and 28% of English word types in a given article are represented in the parallel sentence pairs.

In this work, we seek to learn a joint distribu-

[1] Only 12% of sentences from generally longer English articles have a near-parallel counterpart in the foreign language.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1077–1088,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

tion of translation probabilities over all source and target word pairs in the NEW-domain. We begin with a maximum likelihood estimate of the joint based on a word aligned OLD-domain corpus and update this distribution using NEW-domain comparable data. We define a model based on a single comparable corpus and then extend it to learn from document aligned comparable corpora with any number of comparable *document pairs*. This approach allows us to identify translations for OOV words in the OLD-domain (e.g. French *cisaillement* and *perçage*, which translate as *shear* and *drilling*, in the scientific domain) as well as new translations for previously observed NTS words (e.g. *enceinte* translates as *enclosures*, not *place*, in the scientific domain). In our MT experiments, we use the learned NEW-domain joint distribution to update our SMT model with translations of OOV and low frequency words; we leave the integration of new translations for NTS words to future work.

Our approach crucially depends on finding comparable document pairs relevant to the NEW-domain. Such pairs could be derived from a number of sources, with document pairings inferred from timestamps (e.g. news articles) or topics (inferred or manually labeled). We use Wikipedia[2] as a source of comparable pairs. So-called “interwiki links” (which link Wikipedia articles written on the same topic but in different languages) act as rough guidance that pages may contain similar information. Our approach does not exploit any Wikipedia structure beyond this signal, and thus is portable to alternate sources of comparable articles, such as multilingual news articles covering the same event.

Our model also relies on the assumption that each comparable document pair describes generally the same concepts, though the order and structure of presentation may differ significantly. The efficacy of this method likely depends on the degree of comparability of the data; exploring the correlation between comparability and MT performance is an interesting question for future work.

2 Previous Work

In prior work (Irvine et al., 2013), we presented a systematic analysis of errors that occur when shifting domains in machine translation. That work concludes that errors resulting from unseen (OOV) and new translation sense words cause the majority of the degradation in translation performance that occurs when an MT model trained on OLD-domain data is used to translate data in a NEW-domain. Here, we target OOV errors, though our marginal matching method is also applicable to learning translations for NTS words.

A plethora of prior work learns bilingual lexicons from monolingual and comparable corpora with many signals including distributional, temporal, and topic similarity (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Schafer and Yarowsky, 2002; Schafer, 2006; Klementiev and Roth, 2006; Koehn and Knight, 2002; Haghighi et al., 2008; Mimno et al., 2009; Mausam et al., 2010; Prochasson and Fung, 2011; Irvine and Callison-Burch, 2013). However, this prior work stops short of using these lexicons in translation. We augment a baseline MT system with learned translations.

Our approach bears some similarity to Ravi and Knight (2011), Dou and Knight (2012), and Nuhn et al. (2012); we learn a translation distribution despite a lack of parallel data. However, we focus on the domain adaptation setting. Parallel data in an OLD-domain acts as a starting point (prior) for this translation distribution. It is reasonable to assume an initial bilingual dictionary can be obtained even in low resource settings, for example by crowdsourcing (Callison-Burch and Dredze, 2010) or pivoting through related languages (Schafer and Yarowsky, 2002; Nakov and Ng, 2009).

Daumé III and Jagarlamudi (2011) mine translations for high frequency OOV words in NEW-domain text in order to do domain adaptation. Although that work shows significant MT improvements, it is based primarily on distributional similarity, thus making it difficult to learn translations for low frequency source words with sparse word context counts. Additionally, that work reports results using artificially created monolingual corpora taken from separate source and target halves of a NEW-domain parallel corpus, which may have more lexical overlap with the corresponding test set than we could expect from true monolingual corpora. Our work mines NEW-domain-like document pairs from Wikipedia. In this work, we show that, keeping

[2] www.wikipedia.org

data resources constant, our model drastically outperforms this previous approach. Razmara et al. (2013) take a fundamentally different approach and construct a graph using source language monolingual text and identify translations for source language OOV words by pivoting through paraphrases.

Della Pietra et al. (1992) and Federico (1999) explore models for combining foreground and background distributions for the purpose of language modeling, and their approaches are somewhat similar to ours. However, our focus is on translation.

3 Model

Our goal is to recover a probabilistic translation dictionary in a NEW-domain, represented as a joint probability distribution $p^{\mathsf{new}}(s,t)$ over source/target word pairs. At our disposal, we have access to a joint distribution $p^{\mathsf{old}}(s,t)$ from the OLD-domain (computed from word alignments), plus comparable document pairs in the NEW-domain. From these comparable documents, we can extract raw word frequencies on both the source and target side, represented as marginal distributions $q(s)$ and $q(t)$. The key idea is to estimate this NEW-domain joint distribution to be as similar to the OLD-domain distribution as possible, subject to the constraint that its marginals match those of q.

To illustrate our goal, consider an example. Imagine in the OLD-domain parallel data we find that *accorder* translates as *grant* 10 times and as *tune* 1 time. In the NEW-domain comparable data, we find that *accorder* occurs 5 times, but *grant* occurs only once, and *tune* occurs 4 times. Clearly *accorder* no longer translates as *grant* most of the time; perhaps we should shift much of its mass onto the translation *tune* instead. Figure 2 shows the intuition.

First, we present an objective function and set of constraints over joint distributions to minimize the divergence from the OLD-domain distribution while matching both the source and target NEW-domain marginal distributions. Next, we augment the objective with information about word string similarity, which is particularly useful for the French-English language pair. Optimizing this objective with a single pair of source and target marginals can be performed using an off-the-shelf solver. In practice, though, we have a large set of document pairs, each of which can induce a pair of marginals. Using these per-document marginals provides additional information to the learning function but would overwhelm a common solver. Therefore, we present a sequential learning method for approximately matching the large set of document pair marginal distributions. Finally, we describe how we identify comparable document pairs relevant to the NEW-domain.

3.1 Marginal Matching Objective

Given word-aligned parallel data in the OLD-domain and source and target comparable corpora in the NEW-domain, we first estimate a joint distribution $p^{\mathsf{old}}(s,t)$ over word pairs (s,t) in the OLD-domain, where s and t range over source and target language words, respectively. For the OLD-domain joint distribution, we use a simple maximum likelihood estimate based on non-null automatic word alignments (using grow-diag-final GIZA++ alignments (Och and Ney, 2003)). Next, we find source and target marginal distributions, $q(s)$ and $q(t)$, by relative frequency estimates over the source and target comparable corpora. Our goal is to recover a joint distribution $p^{\mathsf{new}}(s,t)$ for the new domain that matches the marginals, $q(s)$ and $q(t)$, but is minimally different from the original joint distribution, $p^{\mathsf{old}}(s,t)$.

We cast this as a linear programming problem:

$$p^{\mathsf{new}} = \arg\min_{p} \left\| p - p^{\mathsf{old}} \right\|_1 \tag{1}$$

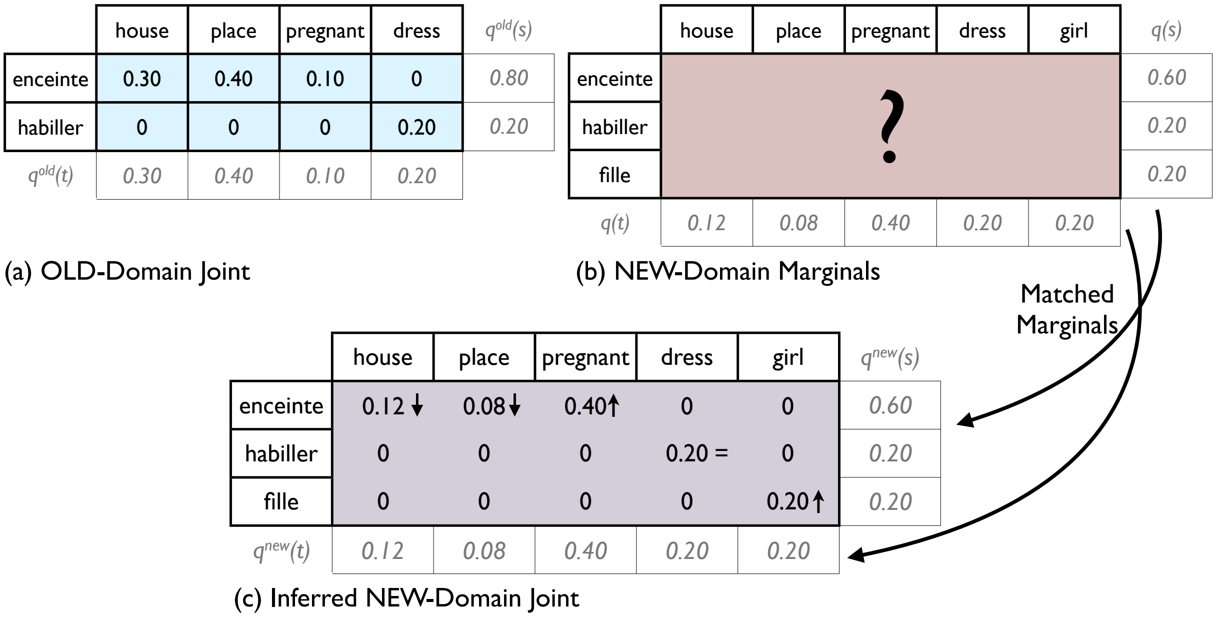
$$\text{subject to: } \sum_{s,t} p(s,t) = 1, \quad p(s,t) \geq 0$$

$$\sum_{s} p(s,t) = q(t), \quad \sum_{t} p(s,t) = q(s)$$

In the objective function, the joint probability matrices p and p^{old} are interpreted as large vectors over all word pairs (s,t). The first two constraints force the result to be a well-formed distribution, and the final two force the marginals to match.

Following prior work (Ravi and Knight, 2011), we would like the matrix to remain as sparse as possible; that is, introduce the smallest number of new translation pairs necessary. A regularization term captures this goal:

$$\Omega(p) = \sum_{\substack{s,t: \\ p^{\mathsf{old}}(s,t)=0}} \lambda_r \times p(s,t) \tag{2}$$



	house	place	pregnant	dress	$q^{old}(s)$
enceinte	0.30	0.40	0.10	0	0.80
habiller	0	0	0	0.20	0.20
$q^{old}(t)$	0.30	0.40	0.10	0.20	

(a) OLD-Domain Joint

	house	place	pregnant	dress	girl	$q(s)$
enceinte	?					0.60
habiller						0.20
fille						0.20
$q(t)$	0.12	0.08	0.40	0.20	0.20	

(b) NEW-Domain Marginals

	house	place	pregnant	dress	girl	$q^{new}(s)$
enceinte	0.12 ↓	0.08 ↓	0.40 ↑	0	0	0.60
habiller	0	0	0	0.20 =	0	0.20
fille	0	0	0	0	0.20 ↑	0.20
$q^{new}(t)$	0.12	0.08	0.40	0.20	0.20	

(c) Inferred NEW-Domain Joint

Figure 2: Starting with a joint distribution derived from OLD-domain data, we infer a NEW-domain joint distribution based on the intuition that the new joint should match the marginals that we observe in NEW-domain comparable corpora. In this example, a translation is learned for the previously OOV word *fille*, and *pregnant* becomes a preferred translation for *enceinte*.

If the old domain joint probability $p^{\text{old}}(s,t)$ was nonzero, there is no penalty. Otherwise, the penalty is λ_r times the new joint probability $p(s,t)$. To discourage the addition of translation pairs that are unnecessary in the new domain, we use a value of λ_r greater than one. Thus, the benefit of a more sparse matrix overwhelms the desire for preventing change. Any value greater than one seems to suffice; we use $\lambda_r = 1.1$ in our experiments.

Inspired by the preference for sparse matrices captured by $\Omega(p)$, we include another orthogonal cue that words are translations of one another: their string similarity. In prior work, string similarity was a valuable signal for inducing translations, particularly for closely related languages such as French and English (Daumé III and Jagarlamudi, 2011). We define a penalty function $f(p)$ as follows: if the normalized Levenshtein edit distance between s *without accents* and t is less than 0.2, no penalty is applied; a penalty of 1 is applied otherwise. We chose the 0.2 threshold manually by inspecting results on our development sets.

$$f(p) = \sum_{s,t} p(s,t) \cdot \begin{cases} 0 & \text{if } \frac{\text{lev}(t,\text{strip}(s))}{\text{len}(s)+\text{len}(t)} < 0.2 \\ 1 & \text{otherwise} \end{cases}$$

The objective function including this penalty is:

$$p^{\text{new}} = \arg\min_{p} \left\| p - p^{\text{old}} \right\|_1 + \Omega(p) + f(p)$$

In principle, additional penalties could be encoded in a similar way.[3] This objective can be optimized by any standard LP solver; we use the Gurobi package (Gurobi Optimization Inc., 2013).

3.2 Document Pair Modification

The above formulation applies whenever we have access to comparable *corpora*. However, often we have access to comparable *documents*, such as those given by Wikipedia inter-language links. We modify our approach to take advantage of the document correspondences within our comparable corpus. In particular, we would like to match the marginals for *all document pairs*.[4] By maintaining separate marginal distributions, our algorithm is presented with more

[3] We experimented with penalties measuring document-pair co-occurrence and monolingual frequency differences but did not see gains on our development sets.

[4] This situation is not unique to our application; multiple marginals are likely to exist in many cases.

information. For example, imagine that one document pair uses “dog” and “chien”, where another document pair uses “cat” and “chat”, each with similar frequency. If we sum these marginals to produce a single marginal distribution, it is now difficult to identify that “dog” should correspond to “chien” and not “chat.” Document pair alignments add information at the cost of additional constraints.

An initial formulation of our problem with multiple comparable document pairs might require the p^{new} marginals to match *all* of the document marginals. In general, this constraint set is likely to result in an infeasible problem. Instead, we take an incremental, online solution, considering a single comparable document pair at a time. For document pair k, we solve the optimization problem in Eq (1) to find the joint distribution minimally different from $p^{\text{k-1}}$, while matching the marginals of *this pair only*. This gives a new joint distribution, tuned specifically for this pair. We then update our current guess of the new domain joint *toward* this document-pair-specific distribution, much like a step in stochastic gradient ascent.

More formally, suppose that before processing the kth document we have a guess at the NEW-domain joint distribution, $p^{\text{new}}_{1:k-1}$ (the subscript indicates that it includes *all* document pairs up to and including document $k-1$). We first solve Eq (1) solely on the basis of this document pair, finding a joint distribution p^{new}_{k} that matches the marginals of the kth document pair *only* and is minimally different from $p^{\text{new}}_{1:k-1}$. Finally, we form a new estimate of the joint distribution by moving $p^{\text{new}}_{1:k-1}$ in the direction of p^{new}_{k}, via:

$$p^{\text{new}}_{1:k} = p^{\text{new}}_{1:k-1} + \eta_u \left[p^{\text{new}}_{k} - p^{\text{new}}_{1:k-1} \right]$$

The learning rate η_u is set to 0.001.[5]

This incremental update of parameters is similar to the margin infused relaxed algorithm (MIRA) (Crammer et al., 2006). Like MIRA and the perceptron, there is not an overall “objective” function that we are attempting to optimize (as one would in many stochastic gradient steps). Instead, we’re aiming for a solution that makes a small amount of progress on each example, in such a way if it received that example again, it would “do better” (in this case: have a closer match of marginals). Also like MIRA, our learning rate is constant. We parallelize learning with mini-batches for increased speed. Eight parallel learners update an initial joint distribution based on 100 document pairs (i.e. each learner makes 100 incremental updates), and then we merge results using an average over the 8 learned joint distributions.

3.3 Comparable Data Selection

It remains to select comparable document pairs. We assume that we have enough monolingual NEW-domain data in one language to rank comparable document pairs (here, Wikipedia pages) according to how NEW-*domain-like* they are. In particular, we estimate the similarity to a source language (here, French) corpus in the NEW domain. For our experiments, we use the French side of a NEW-domain parallel corpus.[6] We could have targeted our learning even more by using our NEW-domain MT test sets. Doing so would increase the chances that our source language words of interest appear in the comparable corpus. However, to avoid overfitting any particular test set, we use the French side of the training data.

For each Wikipedia document pair, we compute the percent of French phrases up to length four that are observed in the French monolingual NEW-domain corpus and rank document pairs by the geometric mean of the four overlap measures. More sophisticated ways to identify NEW-domain-like Wikipedia pages (e.g. Moore and Lewis (2010)) may yield additional performance gains, but, qualitatively, the ranked Wikipedia pages seemed reasonable to the authors.

4 Experimental setup

4.1 Data

We use French-English Hansard parliamentary proceedings[7] as our OLD-domain parallel corpus. With over 8 million parallel lines of text, it is one of the largest freely available parallel corpora for any lan-

[5] We tuned η_u on semi-extrinsic results on the development set. Note that although 0.001 seems small, the values we are moving are *joint probabilities,* which are tiny and so small learning rates make sense.

[6] We could have, analogously, used the *target language* (English) side of the parallel corpus and measure overlap with the English Wikipedia documents, or even used both.

[7] http://www.parl.gc.ca

guage pair. In order to simulate more typical data settings, we sample every 32nd line, using the resulting parallel corpus of $253,387$ lines and $5,051,016$ tokens to train our baseline model.

We test our model using three NEW-domain corpora: (1) the EMEA medical corpus (Tiedemann, 2009), (2) a corpus of scientific abstracts (Carpuat et al., 2013a), and (3) a corpus of translated movie subtitles (Tiedemann, 2009). We use development and test sets to tune and evaluate our MT models. We use the NEW-domain parallel training corpora *only* for language modeling and for identifying NEW-domain-like comparable documents.

4.2 Machine translation

We use the Moses MT framework (Koehn et al., 2007) to build a standard statistical phrase-based MT model using our OLD-domain training data. Using Moses, we extract a phrase table with a phrase limit of five words and estimate the standard set of five feature functions (phrase and lexical translation probabilities in each direction and a constant phrase penalty feature). We also use a standard lexicalized reordering model and two language models based on the English side of the Hansard data and the given NEW-domain training corpora. Features are combined using a log-linear model optimized for BLEU, using the n-best batch MIRA algorithm (Cherry and Foster, 2012). We call this the "simple baseline." In Section 5.2 we describe several other baseline approaches.

4.3 Experiments

For each domain, we use the marginal matching method described in Section 3 to learn a new, domain-adapted joint distribution, $p_k^{\text{new}}(s,t)$, over all French and English words. We use the learned joint to compute conditional probabilities, $p_k^{\text{new}}(t|s)$, for each French word s and rank English translations t accordingly. First, we evaluate the learned joint directly using the distribution based on the word-aligned NEW-domain development set as a gold standard. Then, we perform end-to-end MT experiments. We supplement phrase tables with translations for OOV and low frequency words (we experiment with training data frequencies less than 101, 11, and 1) and include $p_k^{\text{new}}(t|s)$ and $p_k^{\text{new}}(s|t)$ as new translation features for those supplemental translations. For these new phrase pairs, we use the average lexicalized reordering values from the existing reordering tables. For phrase pairs extracted bilingually, we use the bilingually estimated translation probabilities and uniform scores for the new translation features. We experimented with using $p_k^{\text{new}}(t|s)$ and $p_k^{\text{new}}(s|t)$ to estimate additional lexical translation probabilities for the bilingually extracted phrase pairs but did not observe any gains (experimental details omitted due to space constraints). We re-run tuning in all experiments.

We also perform oracle experiments in which we identify translations for French words in word-aligned development and test sets and append these translations to baseline phrase tables.

5 Results

5.1 Semi-extrinsic evaluation

Before doing end-to-end MT experiments, we evaluate our learned joint distribution, $p_k^{\text{new}}(s,t)$, by comparing it to the joint distribution taken from a word aligned NEW-domain parallel development set, $p^{\text{gold}}(s,t)$. We call this evaluation semi-extrinsic because it involves neither end-to-end MT (our extrinsic task) nor an intrinsic evaluation based on our training objective (L1 norm). We find it informative to evaluate the models using bilingual lexicon induction metrics before integrating our output into full MT. That is, we do not compare the full joint distributions, but, rather, for a given French word, how our learned model ranks the word's most probable translation under the gold distribution. In particular, because we are primarily concerned with learning translations for previously unseen words, we evaluate over OOV French word types. In some cases, the correct translation for OOV words is the identical string (e.g. *na*+, *lycium*). Because it is trivial to produce these translations,[8] we evaluate over the subset of OOV development set French words for which the correct translation is not the same string.

Figure 3 shows the mean reciprocal rank for the learned distribution, $p_k^{\text{new}}(s,t)$, for each domains as a function of the number of comparable document pairs used in learning. In all domains, the comparable document pairs are sorted according to their sim-

[8] And, indeed, by default our decoder copies OOV strings into its output directly.

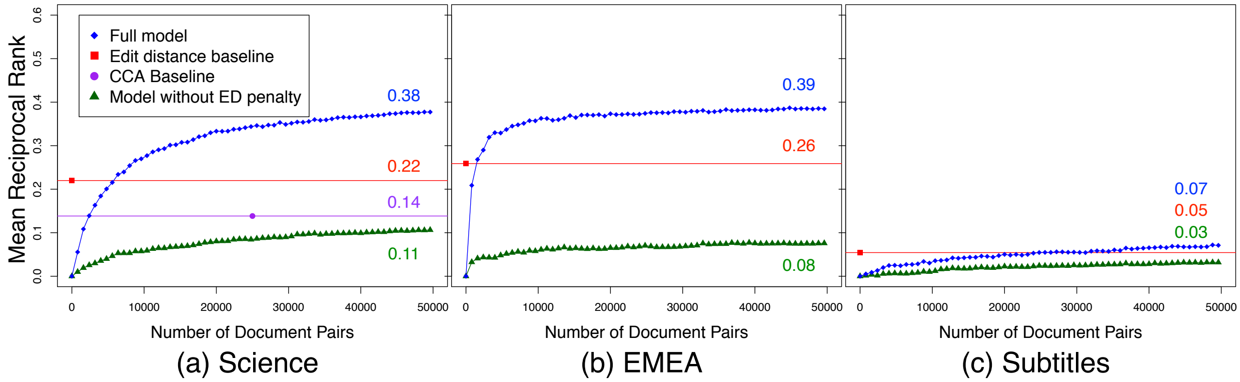


Figure 3: Semi-extrinsic bilingual lexicon induction results. Mean reciprocal rank is computed over all OOV development set words for which identity is not the correct translation.

ilarity with the NEW-domain. Figure 3 also shows the performance of baseline models and our learner without the edit distance penalty. For each source word s, the edit distance (ED) baseline ranks all English words t in our monolingual data by their edit distance with s.[9] The Canonical Correlation Analysis (CCA) baseline uses the approach of Daumé III and Jagarlamudi (2011) and the top $25,000$ ranked document pairs as a comparable corpus. That model performs poorly largely because of sparse word context counts. Interestingly, for Science and EMEA, the performance of our full model at $50,000$ document pairs is higher than the sum of the edit distance baseline and the model without the edit distance penalty, indicating that our approach effectively combines the marginal matching and edit distance signals.

The learning curves for the three domains vary substantially. For Science, learning is gradual and it appears that additional gains could be made by iterating over even more document pairs. In contrast, the model learns quickly for the EMEA domain; performance is stable after $20,000$ document pairs. Given these results and our experience with the two domains, we hypothesize that the difference is due to the fact that the Science data is much more heterogenous than the EMEA data. The Science data includes physics, chemistry, and biology abstracts, among others. The drug labels that make up most of the EMEA data are more homogeneous. In Section 6 we comment on the poor Subtitles performance, which persists in our MT experiments.

We experimented with making multiple learning passes over the document pairs and observed relatively small gains from doing so. In all experiments, learning from some number of additional new document pairs resulted in higher semi-extrinsic performance gains than passing over document pairs which were already observed.

In the case of OOV words, it's clear that learning something about how to translate a previously unobserved French word is beneficial. However, our learning method also learns domain-specific new-translation senses (NTS). Table 1 shows some examples of what the marginal matching method learns for different types of source words (OOVs, low frequency, and NTS).

5.2 MT evaluation

By default, the Moses decoder copies OOV words directly into its translated output. In some cases, this is correct (e.g. *ensembles*, *blumeria*, *google*). In other cases, French words can be translated into English correctly by simply stripping accent marks off of the OOV word and then copying it to the output (e.g. *caméra*, *éléments*, *molécules*). In the Science and EMEA domains, we found that our baseline BLEU scores improved from 21.91 to 22.20 and 23.67 to 24.45, respectively, when we changed the default handling of OOVs to strip accents before

[9] In particular, for each domain and each OOV French word, we ranked the set of all English words that appeared at least five times in the set of 50,000 most NEW-domain like Wikipedia pages. Using a frequency threshold of five helped eliminate French words and improperly tokenized English words from the set of candidates.

French	OLD top $p^{\text{old}}(t\|s)$	NEW top $p^{\text{gold}}(t\|s)$	MM-learned top $p^{\text{new}}(t\|s)$
OOV words			
cisaillement courbure	- -	shear strength shearing curvature bending curvatures	shear viscous newtonian curvature curved manifold
Low frequency words			
linéaires récepteur	linear receiver	linear nonlinear non-linear receptor receiver y1	linear linearly nonlinear receptor receiver receptors
New translation sense words			
champ marche	field jurisdiction scope working march work	field magnetic near-field walk step walking	field magnetic fields march walk walking

Table 1: Hand-picked examples of Science-domain French words and their top English translations in the OLD-domain, NEW-domain, and marginal matching distributions. The first two are OOVs. The next two only appeared four and one time, respectively, in the training data and only aligned to a single English word. The last two are NTS French words: words that appeared frequently in the training data but for which the word's sense in the new domain shifts.

copying into the output. Interestingly, performance on the Subtitles domain text did not change at all with this baseline modification. This is likely due to the fact that there are fewer technical OOVs (the terms typically captured by this accent-stripping pattern) in the subtitles domain.

Throughout our experiments, we found it critical to retain correct 'freebie' OOV translations. In the results presented below, including the baselines, we supplement phrase tables with a new candidate translation but also include accent-stripped identity, or 'freebie,' translations in the table for all OOV words. We experimented with classifying French words as freebies or needing a new translation, but oracle experiments showed very little improvement (about 0.2 BLEU improvement in the Science domain), so instead we simply include both types of translations in the phrase tables.

In addition to the strip-accents baseline, we compare results with four other baselines. First, we drop OOVs from the output translations. Second, like our semi-extrinsic baseline, we rank English words by their edit distance away from each French OOV word (ED baseline). Third, we rank English words by their document-pair co-occurrence score with each French OOV word. That is, for all words w, we compute $D(w)$, the vector indicating the document pairs in which w occurs, over the set of 50,000 document-pairs which are most NEW-*domain-like*. For French and English words s and t, if $D(s)$ and $D(t)$ are dissimilar, it is less likely (s, t) is a valid translation pair. We weight $D(w)$ entries with BM25 (Robertson et al., 1994). For all French OOVs, we rank all English translations according to the cosine similarity between the pair of $D(w)$ vectors. The fourth baseline uses the CCA model described in Daumé III and Jagarlamudi (2011) to rank English words according to their distributional similarity with each French word. For the CCA baseline comparison, we only learned translations using 25,000 Science-domain document pairs, rather than the full 50,000 and for all domains. However, it's unlikely that learning over more data would overcome the low performance observed so far. For the final three baselines, we append French OOV words and their highest ranked English translation to the phrase table. Along with each new translation pair, we include one new phrase table feature with the relevant translation score (edit distance, document similarity, or CCA distributional similarity). For all baselines other than drop-OOVs, we also include accent-stripped translation pairs with an additional indicator feature.

Table 3 shows results appending the top ranked English translation for each OOV French word using each baseline method. None of the alternate baselines outperform the simplest baseline on the subtitles data. Using document pair co-occurrences is the strongest baseline for the Science and EMEA domains. This confirms our intuition that taking advantage of document pair alignments is worthwhile. For Science and EMEA, supplementing a model with OOV translations learned through our marginal matching method drastically outperforms all base-

OOVs translated correctly and incorrectly	
Input	les résistances au **cisaillement** par **poinçonnement** ...
Ref	the punching **shear strengths**...
Baseline	the resistances in **cisaillement** by **poinconnement** ...
MM	the resistances in **shear reinforcement**...

OOV translated incorrectly	
Input	présentation d' un logiciel permettant de gérer les données **temporelles** .
Ref	presentation of software which makes it possible to manage **temporal** data .
Baseline	introduction of a software to manage **temporelles** data .
MM	introduction of a software to manage data **plugged** .

Low frequency French words	
Input	...limite est liée à la **décroissance** très rapide du **couplage** électron-phonon avec la température .
Ref	...limit is linked to the rapid **decrease** of the electron-phonon **coupling** with temperature .
Baseline	...limit is linked to the **decline** very rapid electron-phonon **linkage** with the temperature .
MM	...limit is linked to the **linear** very rapid electron-phonon **coupling** with the temperature .

Table 2: Example MT outputs for Science domain. The baseline strips accents (Table 3). In the first example, the previously OOV word *cisaillement* is translated correctly by an MM-supplemented model. The OOV *poinçonnement* is translated as *reinforcement* instead of *strengths*, which is incorrect with respect to the reference but arguably not bad. In the second example, *temporelles* is not translated correctly in the MM output. In the third example, the MM-hypothesized correct translation of low frequency word *couplage*, *coupling*, is chosen instead of incorrect *linkage*. Also in the third example, the low frequency word *décroissance* is translated as the MM-hypothesized incorrect translation *linear*. In the case of *décroissance*, the baseline's translation, *decline*, is much better than the MM translation *linear*.

lines. Using our model to translate OOV words yields scores of 23.62 and 26.97 in the Science and EMEA domains, or 1.19 and 1.94 BLEU points, respectively, above the strongest baseline. We observe additional gains by also supplementing the model with translations for low frequency French words. For example, when we use our approach to translate source words in the Science domain which appear ten or fewer times in our OLD-domain training data, the BLEU score increases to 24.28.

We tried appending top-k translations, varying k. However, we found that for the baselines as well as our MM translations, using only the top-1 English translations outperformed using more.

Table 3 also shows the result of supplementing a baseline phrase table with oracle OOV translations. Using the marginal matching learned OOV translations takes us 30% and 40% of the way from the baseline to the oracle upper bound for Science and EMEA, respectively.

We have focused on supplementing an SMT model trained on a sample of the Hansard parallel corpus in order to mimic typical data conditions, but we have also performed experiments supplementing a model trained on the full dataset.[10] Beginning with the larger model, we observe performance gains of 0.8 BLEU points for both the EMEA and the Science domains over the strongest baselines, which are based on document similarity, when we add OOV

	Science	EMEA	Subs
Simple Baseline	21.91	23.67	**13.18**
Drop OOVs	20.22	18.95	11.86
Accent-Stripped	22.20	24.45	13.13
ED Baseline	22.10	24.35	12.95
Doc Sim Baseline.	**22.43**	**25.03**	13.02
CCA Baseline	21.41	-	-
MM Freq<1 (OOV)	23.62	26.97	**13.07**
MM Freq<11	**24.28**	**27.26**	12.97
MM Freq<101	23.96	26.82	12.92
Oracle OOV	26.38	29.99	15.06

Table 3: BLEU results using: (1) baselines, (2) phrase tables augmented with top-1 translations for French words with indicated OLD training data frequencies, (3) phrase tables augmented with OOV oracle translations.

[10] We still use the joint that was learned starting with the one estimated over the sample; we may observe greater gains over the full Hansard baseline with a stronger initial joint.

translations. As expected, these gains are less than what we observe when our baseline model is estimated over less data, but they are still substantial.

In all experiments, we have assumed that we have no NEW-domain parallel training data, which is the case for the vast majority of language pairs and domains. However, In the case that we do have some NEW-domain parallel data, OOV rates will be somewhat lower, but our method is still applicable. For example, we would need 2.3 million words of Science (NEW-domain) parallel data to cover just 50% of the OOVs in our Science test set, and 4.3 million words to cover 70%.

6 Discussion

BLEU score performance gains are substantial for the Science and EMEA domains, but we don't observe gains on the subtitles text. We believe this difference relates to the difference between a corpus domain and a corpus register. As Lee (2002) explains, a text's *domain* is most related to its topic, while a text's *register* is related to its type and purpose. For example, religious, scientific, and dialogue texts may be classified as separate registers, while political and scientific expositions may have a single register but different domains. Our science and EMEA corpora are certainly different in domain from the OLD-domain parliamentary proceedings, and our success in boosting MT performance with our methods indicates that the Wikipedia comparable corpora that we mined match those domains well. In contrast, the subtitles data differs from the OLD-domain parliamentary proceedings in both domain and register. Although the Wikipedia data that we mined may be closer in domain to the subtitles data than the parliamentary proceedings,[11] its register is certainly not film dialogues.

Although the use of marginal matching is, to the best of our knowledge, novel in MT, there are related threads of research that might inspire future work. The intuition that we should match marginal distributions is similar to work using no example labels but only label proportions to estimate labels, for example in Quadrianto et al. (2008). Unlike that work, our label set corresponds to entire vocabularies, and we have multiple observed label proportions. Also, while the marginal matching objective seems effective in practice, it is difficult to optimize. A number of recently developed approximate inference methods use a decomposition that bears a strong resemblance to this objective function. Considering the marginal distributions from each document pair to be a separate subproblem, we could approach the global objective of satisfying all subproblems as an instance of dual decomposition (Sontag et al., 2010) or ADMM (Gabay and Mercier, 1976; Glowinski and Marrocco, 1975).

We experiment with French-English because tuning and test sets are available in several domains for that language pair. However, our techniques are directly applicable to other language pairs, including those that are less related. We have observed that many domain-specific terms, particularly in medical and science domains, are borrowed across languages, whether or not the languages are related. Even for languages with different character sets, one could do transliteration before measuring orthographical similarity.

Although we were able to identify translations for some NTS words (Table 1), we did not make use of them in our MT experiments. Recent work has identified NTS words in NEW-domain corpora (Carpuat et al., 2013b), and in future work we plan to incorporate discovered translations for such words into MT.

7 Conclusions

We proposed a model for learning a joint distribution of source-target word pairs based on the idea that its marginals should match those observed in NEW-domain comparable corpora. Supplementing a baseline phrase-based SMT model with learned translations results in BLEU score gains of about two points in the medical and science domains.

Acknowledgments

We gratefully acknowledge the support of the 2012 JHU Summer Workshop and NSF Grant No 1005411. We would like to thank the entire DAMT team (`http://hal3.name/damt/`) and Sanjeev Khudanpur for their help and suggestions. We also acknowledge partial support from DARPA

[11] In fact, we believe that it is. Wikipedia pages that ranked very high in our subtitles-like list included, for example, the movie *The Other Side of Heaven* and actor *Frank Sutton*.

CSSG Grant D11AP00279 and DARPA BOLT Contract HR0011-12-C-0015 for Hal Daumé III and support from the Johns Hopkins University Human Language Technology Center of Excellence for Ann Irvine. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. 2013a. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013b. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, December.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

S. Della Pietra, V. Della Pietra, R. L. Mercer, and S. Roukos. 1992. Adaptive language modeling using minimum discriminant estimation. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marcello Federico. 1999. Efficient language model adaptation through mdi estimation. In *Proceedings of EUROSPEECH*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Daniel Gabay and Bertrand Mercier. 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1):17 – 40.

Roland Glowinski and A. Marrocco. 1975. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de dirichlet non linéaires. *Rev. Franc. Automat. Inform. Rech. Operat.*, 140:41–76.

Gurobi Optimization Inc. 2013. Gurobi optimizer reference manual.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

David Lee. 2002. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language and Computers*, 42(1):247–292.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, and Jeff Bilmes. 2010. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174:619–637, June.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, and Quoc V. Le. 2008. Estimating labels from label proportions. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Majid Razmara, Maryam Siahbani, Gholamreza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Text REtrieval Conference*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Charles Schafer. 2006. *Translation Discovery Using Diverse Similarity Measures*. Ph.D. thesis, Johns Hopkins University.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

David Sontag, A. Globerson, and Tommi Jaakola, 2010. *Introduction to dual decomposition for inference*, chapter 1. MIT Press.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*.