Abstract

We use bilingual lexicon induction techniques, which learn translations from monolingual texts in two languages, to build an end-to-end statistical machine translation (SMT) system without the use of any bilingual sentence-aligned parallel corpora. We present detailed analysis of the accuracy of bilingual lexicon induction, and show how a discriminative model can be used to combine various signals of translation equivalence (like contextual similarity, temporal similarity, orthographic similarity and topic similarity). Our discriminative model produces higher accuracy translations than previous bilingual lexicon induction techniques. We reuse these signals of translation equivalence as features on a phrase-based SMT system. These monolingually-estimated features enhance low resource SMT systems in addition to allowing end-to-end machine translation without parallel corpora.

Natural Language Engineering **1** (1): 1–34. Printed in the United Kingdom

End-to-End Statistical Machine Translation with Zero or Small Parallel Texts

Ann Irvine and Chris Callison-Burch

(*Received December 15, 2014*)

1 Introduction

SMT typically relies on very large amounts of bilingual sentence-aligned parallel texts. Here, we consider settings in which we have access to (1) bilingual dictionaries but no parallel sentences for training, and (2) only a small amount of parallel training data. In the first case, we augment a baseline system that produces a simple dictionary gloss with additional translations that are learned using monolingual corpora in the source and target languages. In the second case, we wish to augment a baseline statistical model learned over small amounts of parallel training data with additional translations and features estimated over monolingual corpora.

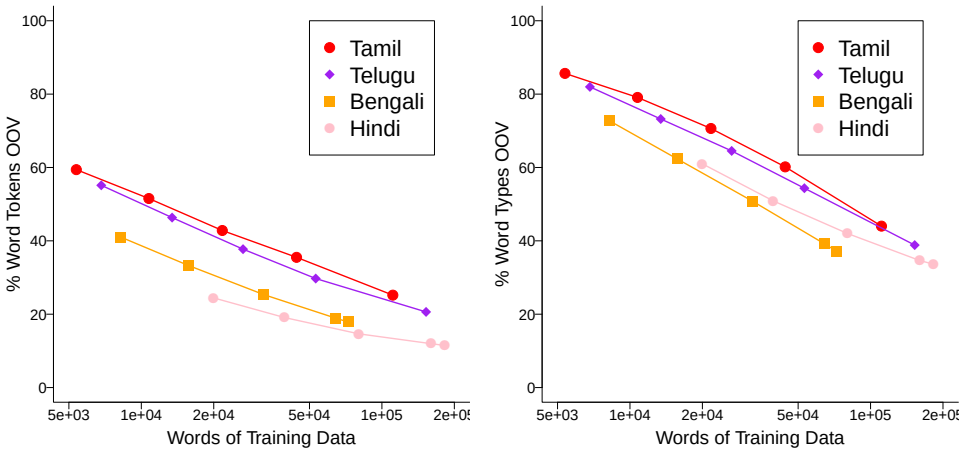
In this article, we detail our approach to bilingual lexicon induction, which allows us to learn translations from independent monolingual texts or comparable corpora that are written in two languages (Section 2). We evaluate the accuracy of our model on correctly learning dictionary translations, and examine its performance on low frequency words which are more likely to be out of vocabulary (OOV) with respect to the training data for SMT systems.

We describe our approach to learning how to transliteration from one language's script into another language's script (Section 3). Transliteration is a useful aid, since many OOV items correspond to named entities or technical terms, which are often transliterated rather than translated.

We show how the diverse signals of translation equivalence that we use in our discriminative model for bilingual lexicon induction can also be used as additional features for a phrase table in a standard SMT model to enhance low resource SMT systems (Section 4). We analyze 6 low resource languages and find consistent improvements in BLEU score when we incorporate translations of OOV items and when we re-score the phrase table with additional monolingually estimated feature functions.

Finally, we combine all of these ideas and demonstrate how to build a true end-to-end SMT system without bilingual sentence-aligned parallel corpora (Section 5). We build a patchwork phrase table out of entries from a standard bilingual dictionaries, plus induced translations, plus transliterations. We associate each translation with a set of monolingually-estimated feature functions and generate translations using a SMT decoder that incorporates these scores and a language model probability.

This article combines and extends several of our past papers on this topic: (Irvine,



(a) Tokens (b) Types

Fig. 1: The rate of out of vocabulary (OOV) items for six low resources languages. We show the token-based and type-based OOV rates. The curves are generated by randomly sampling the training datasets described in Section 4.1.

Callison-Burch, and Klementiev2010), (Irvine and Callison-Burch2013b), (Irvine and Callison-Burch2013a), (Irvine2014) and (Irvine and Callison-BurchIn submission). This article expands the previous publications by providing additional analysis and examples from Ann Irvine's PhD thesis. The main experimental results that were not previously published are the expanded set of experiments on our discriminative model for bilingual lexicon induction (Section 2). Because this article assembles research undertaken over a period of 5+ years, it is not perfectly consistent from section to section in terms of what languages it analyzes or in using identical features across all experiments. Despite this, we believe that this article provides a valuable synthesis of our past work on trying to improve SMT for low resource languages, with the aim of reducing or eliminating the dependency on sentence-aligned bilingual parallel corpora.

2 Learning Translations of Unseen Words

SMT typically uses sentence-aligned bilingual parallel texts to learn the translations of individual words (Brown et al.1990). Another thread of research has examined *bilingual lexicon induction* which tries to induce translations from monolingual corpora in two languages. These monolingual corpora can range from being completely unrelated topics to being comparable corpora. Here we examine the usefulness of bilingual lexicon induction as a way of augmenting SMT when we only have access to small bilingual parallel corpora, and when we have no bitexts whatsoever.

The most prominent problem that arises when a machine translation system has access to limited parallel resources is the fact that there are many unknown words

that are OOV with respect to the training data, but which do appear in the texts that we would like the SMT system to translate. Figure 1 quantifies the rate of OOVs for half a dozen low resource languages. It shows the percent of word tokens and word types in a development set that are OOV with respect to varying amounts of training data for several Indian languages.[1] Bilingual lexicon induction can be used to try to improve the coverage of our low resource translation models, by learning the translations of words that do not occur in the parallel training data.

Although past research into bilingual lexicon induction has been motivated by the idea that it could be used to improve machine translation systems by translating OOV words, it has rarely been evaluated that way. Notable exceptions of past research that does evaluate bilingual lexicon induction in the context of machine translation through better OOV handling include (Daumé and Jagarlamudi2011), (Dou and Knight2013) and (Dou, Vaswani, and Knight2014). However, the majority of prior work in bilingual lexicon induction has treated it as a standalone task, without actually integrating induced translations into end-to-end machine translation. It was instead evaluated by holding out a portion of a bilingual dictionary and evaluating how well the algorithm learns the translations of the held out words. In this article, we perform a systematic examination of the efficacy of bilingual lexicon induction for end-to-end translation.

Bilingual lexicon induction uses monolingual or comparable corpora, usually paired with a small seed dictionary, to compute signals of translation equivalence. Here we briefly describe our approach to bilingual lexicon induction that combines multiple signals of translation equivalence in a discriminative model. More details about our approach are available in (Irvine and Callison-Burch2013b), (Irvine2014), and (Irvine and Callison-BurchIn submission). Although past research into bilingual lexicon induction also explored multiple signals of translation equivalence (for instance, (Schafer and Yarowsky2002)), these features have not previously been combined using a discriminative model.

2.1 Our approach to bilingual lexicon induction

We frame bilingual lexicon induction as a binary classification problem: for a pair of source and target language words, we predict whether the two are translations of one another or not. Since binary classification does not inherently give us a list of the best translations, we need to take an additional step. For a given source language word we find its best translation or its n-best translations by first using our classifier on all target language words. We then rank them based on how confident the classifier is that each target word is a translation of the source word. The features used by our classifier include a variety of signals of translation equivalence that are drawn from past work in bilingual lexicon induction, notably by (Rapp1995;

[1] Our Indian language datasets are described in Section 4.1. Note that in this OOV analysis, we do not include the dictionaries, only complete sentences of bilingual training data.

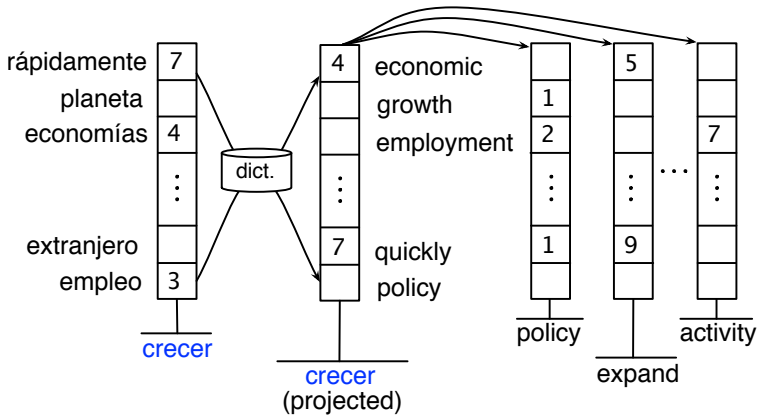


Fig. 2: Example of projecting contextual vectors over a seed bilingual lexicon. The Spanish word *crecer* appears in the context of the words *empleo, extranjero*, etc in monolingual texts. We use this co-occurence information to build a context vector. Each position in the context vector for corresponds to a word in the Spanish vocabulary. The vector for *crecer* is projected into the English vector space using a small seed dictionary. Context vectors for all English words (*policy, expand*, etc.) are collected and then compared against the projected context vector for Spanish *crecer*. Finally, *contextual similarities* are calculated by comparing the projected vector with the context vector of each target word using cosine similarity. Word pairs with high cosine similarity are likely to be translations of one another.

Fung1995; Schafer and Yarowsky2002; Klementiev and Roth2006; Klementiev et al.2012), and others. The features that we use in our model are:

- **Contextual similarity** – In a similar fashion to how vector space models can be used to compute the similarity between two words in one language by creating vectors that representing their co-occurrence patterns with other words (Turney and Pantel2010), context vector representations can also be used to compare the similarity of words across two languages. The earliest work in bilingual lexicon induction by (Rapp1995) and (Fung1995) used the surrounding context of a given word as a clue to its translation. (Fung and Yee1998) and (Rapp1999), used small seed dictionaries to *project word-based context vectors* from the vector space of one language into the vector space of the other language. We use the vector space approach of (Rapp1999) to compute similarity between word in the source and target languages.
  More formally, assume that $(s_1, s_2, \ldots s_N)$ and $(t_1, t_2, \ldots t_M)$ are (arbitrarily indexed) source and target vocabularies, respectively. A source word f is represented with an N-dimensional vector and a target word e is represented with an M-dimensional vector (see Figure 2). The component values of the vector representing a word correspond to how often each of the words in that vocabulary appear within a two word window on either side of the given word. These counts are collected using monolingual corpora. After the values have

been computed, a contextual vector for f is projected onto the English vector space using translations in a given bilingual dictionary to map the component values into their appropriate English vector positions. This sparse projected vector is compared to the vectors representing all English words, e. Each word pair is assigned a contextual similarity score based on the similarity between e and the projection of f.

Various means of computing the component values and vector similarity measures have been proposed in literature (e.g. (Fung and Yee1998; Rapp1999)). Following (Fung and Yee1998), we compute the value of the k-th component of f's contextual vector, f_k, as follows:

$$f_k = n_{f,k} * (log(n/n_k) + 1) \quad (1)$$

where $n_{f,k}$ and n_k are the number of times s_k appears in the context of f and in the entire corpus, and n is the maximum number of occurrences of any word in the data. Intuitively, the more frequently s_k appears with f_i and the less common it is in the corpus in general, the higher its component value. After projecting each component of the source language contextual vectors into the English vector space, we are left with M-dimensional source word contextual vectors, $F_{context}$, and correspondingly ordered M-dimensional target word contextual vectors, $E_{context}$, for all words in the vocabulary of each language. We use cosine similarity to measure the similarity between each pair of contextual vectors:

$$sim_{context}(F_{context}, E_{context}) = \frac{F_{context} \cdot E_{context}}{||F_{context}||||E_{context}||} \quad (2)$$

- **Temporal similarity** – Usage of words over time may be another signal of translation equivalence. The intuition that is that news stories in different languages will tend to discuss the same world events on the same day and, correspondingly, we expect that source and target language words which are translations of one another will appear with similar frequencies over time in monolingual data. For instance, if the English word *tsunami* is used frequently during a particular time span, the Spanish translation *maremoto* is likely to also be used frequently during that time. To calculate temporal similarity, we collected online monolingual newswire over a multi-year period and associate each article with a time stamps. We gather temporal signatures for each source and target language unigram from our time-stamped web crawl data in order to measure temporal similarity, in a similar fashion to (Schafer and Yarowsky2002; Klementiev and Roth2006; Alfonseca, Ciaramita, and Hall2009). We calculate the temporal similarity between a pair of words, using the method defined by (Klementiev and Roth2006).
- **Orthographic similarity** – Words that are spelled similarly are sometimes good translations, since they may be etymologically related, or borrowed words, or the names of people and places. We compute the orthographic

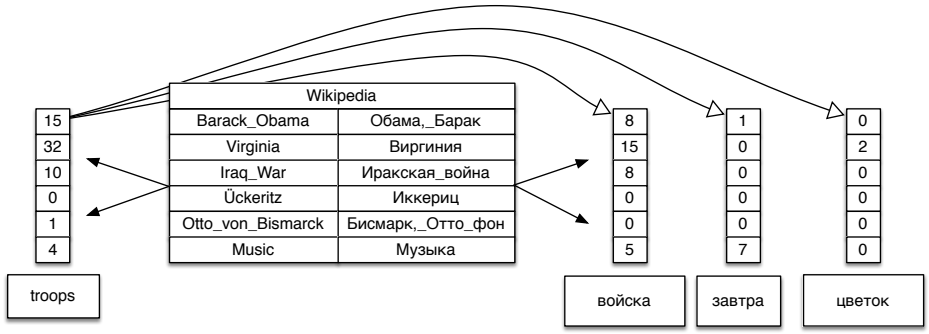


Fig. 3: Illustration of how we compute the topical similarity between *troops* and three Russian candidate translations. We first collect the topical signatures for each word (e.g. *troops* appears in the page about *Barack Obama* 15 times and in the page about *Virginia* 32 times.) based on the interlingually linked pages. We can then directly compare each pair of topical signatures.

similarity between a pair of words using Levenshtein edit distance[2], normalized by the average of the lengths of the two words. This is straightforward for languages which use the same character set, but it is more complicated for languages that are written using different scripts. For non-Roman script languages, we transliterate words into the Roman script before measuring orthographic similarity with their candidate English translations (Virga and Khudanpur2003; Irvine, Callison-Burch, and Klementiev2010). More details of our transliteration method are given in Section 3.

- **Topic similarity** – Articles that are written about the same topic in two languages, are likely to contain words and their translations, even if the articles themselves are written independently and are not translations of one another. We use Wikipedia's interlingual links to identify comparable articles across languages. These links define a number of topics, and we construct a topic vector. We compute cosine distance between topic signatures.

$$sim_{topic}(F_{topic}, E_{topic}) = \frac{F_{topic} \cdot E_{topic}}{||F_{topic}||||E_{topic}||}, \tag{3}$$

The length of a word's topic vector is the number of interlingually linked article pairs. Each component f_k of F_{topic} is the count of the word f in the foreign article from the kth linked article pair, normalized by the total occurrences of k. The dimensionality of the topic signatures varies depending on the language pair. The number of linked articles in Wikipedia range from 84 (between Kashmiri and English) to over 500 thousand (between French and English). Figure 3 illustrates this signal. More details on our topic similarity are in (Irvine2014).

[2] `http://en.wikipedia.org/wiki/Levenshtein_distance`

- **Frequency similarity** – Words that are translations of one another are likely to have similar relative frequencies in monolingual corpora. We measure the frequency similarity of two words, sim_{freq}, as the absolute value of the difference between the log of their relative corpus frequencies, or:

$$sim_{freq}(e, f) = |log(\frac{freq(e)}{\sum_i freq(e_i)}) - log(\frac{freq(f)}{\sum_i freq(f_i)})| \quad (4)$$

  This helps prevent high frequency closed class words from being considered viable translations of less frequent open class words.
- **Burstiness similarity** – Burstiness is a measure of how peaked a word's usage is over a particular corpus of documents (Pierrehumbert2012). Bursty words are topical words that tend to appear when some topic is discussed in a document. For example, *earthquake* and *election* are considered bursty. In contrast, non-bursty words are those that appear more consistently throughout documents discussing different topics, *use* and *they*, for example. (Church and Gale1995; Church and Gale1999) provide an overview of several ways to measure burstiness empirically. Following (Schafer and Yarowsky2002), we measure the burstiness of a given word based on Inverse Document Frequency (IDF):

$$IDF_w = -log\frac{df_w}{|D|}, \quad (5)$$

  where df_w is the number of documents that w appears in, and $|D|$ is the total number of documents in the collection. We have also experimented with a second burstiness measure, similar to that defined by (Church and Gale1995), as the average frequency of w divided by the percent of documents in which w appears. We make one modification to the definition provided by (Church and Gale1995) and use relative frequencies rather than absolute frequencies to account for varying document lengths:

$$B_w = \frac{\sum_{d_i \in D} rf_{w_{d_i}}}{df_w}, \quad (6)$$

  where, as before, df_w is the number of documents in which w appears and $rf_{w_{d_i}}$ is the relative frequency of w in document d_i. Relative frequencies are raw frequencies normalized by document length.
- We also compute a number of variations on the above using word prefixes and suffixes instead of fully inflected words, and based on two different sources of data (web crawls and Wikipedia). In total, our model uses 18 such features in order to rank English words as potential translations of the input foreign word.

Table 1 shows some examples of the highest ranking English translations of 5 Spanish words for several of our signals of translation equivalence. Each signal produces different types of errors. For instance, using topic similarity, *montana*, *miley*, and *hannah* are ranked highly as candidate translations of the Spanish word *montana*. The TV character Hannah Montana is played by actress Miley Cyrus, so the topic similarity between these words makes sense.

alcanzaron	*sanitario*	*desarrollos*	*volcánica*	*montana*
CONTEXTUAL SIMILARITY				
reached	exil	advances	**volcanic**	arendt
enjoyed	rhombohedral	**developments**	eruptive	**montana**
contained	apt	changes	coney	glasse
contains	immune	placing	rhonde	teter
TEMPORAL SIMILARITY				
travel	snowpocalypse	occupied	wawel	dzv
road	airport	aer	**volcanic**	spatz
news	dioxide	madoff	ash	centimes
services	steinmeier	declaration	spewed	kleve
ORTHOGRAPHIC SIMILARITY				
alcantara	**sanitary**	ferroalloy	**volcanic**	**montana**
albanian	sanitation	barrosos	volcanism	fontana
lazzaroni	unitario	destroyers	voltaic	montane
lanaro	sanitarium	mccarroll	vacancy	mentana
TOPIC SIMILARITY				
reached	health	**developments**	**volcanic**	**montana**
began	transcultural	developed	eruptions	miley
led	medical	development	volcanism	hannah
however	sanitation	used	lava	beartooth

Table 1: Examples of translation candidates ranked using contextual similarity, temporal similarity, orthographic similarity and topic similarity. The correct English translations, when found, are bolded.

A significant research challenge is how best to combine these signals. Previous approaches have combined signals in an unsupervised fashion. One method of combining the ranked lists of translations that are independently generated by each of the signals of translation equivalence is using mean reciprocal rank (MRR), which is a measure typically used in information retrieval. It is defined as the average of

Language	Dict entries (freq >= 10)	Wikipedia words	interlanguage links	Web crawl words	Web crawl dates
Bengali	5,368	4,998,454	18,603	8,295,164	467
Hindi	6,585	16,198,183	25,078	31,123,091	823
Tamil	4,735	9,154,660	23,468	3,928,554	157
Telugu	5,136	8,769,259	8,841	3,254,373	120

Table 2: Statistics about the data used in our bilingual lexicon induction experiments.

the reciprocal ranks of results for a sample of queries Q:[3]

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{7}$$

In the case of bilingual lexicon induction we query each signal of translation equivalence with a source word, the value $|Q|$ corresponds to the number of signals, and rank_i corresponds to the rank of a target language translation under the i^{th} signal. The translation with the highest MRR value is output as the best translation. The disparate of signals of translation equivalence all provide an equal contribution in MRR, regardless of how good they are at picking out good translations.

Instead of weighting each signal equally, we use a discriminative model that is trained using entries in the seed bilingual dictionary as positive examples of translations, and random word pairs as negative examples (we use a 1:3 ratio of positive to negative examples). Discriminative models have an advantage over MRR in that they are able to weight the contribution of each feature based on how well it predicts the translations of words in a development set. When feature weights are discriminatively set, these signals produce dramatically higher translation quality than MRR. In (Irvine and Callison-BurchIn submission) we present experimental results showing consistent improvements in translation accuracy for 25 languages. The absolute accuracy increases over the MRR baseline ranges from 5%-31%, which correspond to 36%-216% relative improvements. Our discriminative approach requires a small number of translations to use as a development set. This requirement is not a major imposition, since bilingual lexicon induction already typically requires a small seed bilingual dictionary.

2.2 Experiments with bilingual lexicon induction

We excerpt a number of experiments from (Irvine and Callison-BurchIn submission) that show our method's performance on four of the Indian languages that we examine in the end-to-end machine translation experiments (Section 5).

[3] `http://en.wikipedia.org/wiki/Mean_reciprocal_rank`

Data We created bilingual dictionaries using native-language informants on Amazon Mechanical Turk (MTurk). In (Pavlick et al.2014), we describe a study of the languages demographics of workers on MTurk. In that work, we focused on the 100 languages which have the largest number of Wikipedia articles and posted Human Intelligence Tasks (HITs) asking workers to translate the $10,000$ most frequent words in the $1,000$ most viewed pages for each source language. For the experiments in this article, we filter the dictionaries to include only high quality translations. Specifically, we limit ourselves to words that occurred at least 10 times in our monolingual data sets, and we only use translations that have a quality score of at least 0.6 under the worker quality metric defined by (Pavlick et al.2014). Workers provided between 1–32 reference translations for each word (with an average of 1.4 translations per word).

We gathered monolingual data sets by scraping online newspapers in each language, and by downloading the content of each language version of Wikipedia. For all languages, we use Wikipedia's January 2014 data snapshots. Table 2 gives statistics about the monolingual data sets.

Measuring accuracy We measure performance using accuracy in the top-k ranked translations. We define top-k accuracy over some set of ranked lists L as follows:

$$acc_k = \frac{\sum_{l \in L} I_{lk}}{|L|} \tag{8}$$

where I_{lk} is an indicator function that is 1 if and only if a correct item is included in the top-k elements of list l. That is, top-k accuracy is the proportion of ranked lists in a set of ranked lists for which a correct item is included anywhere in the highest k ranked elements. The denominator $|L|$ is the number of words in a test set for a language. The numerator indicates how many of the words had at least one correct translation in the top-k translations posited for the word. Top-k accuracy increases as k increases.

A translation counts as correct if it appears in our bilingual dictionary for the language. We split our dictionaries into separate training and test sets. The test sets consist of $1,000$ randomly selected source language words and their translations. The training sets consist of the remaining words. We use the training set to project vectors for contextual similarity, and to train the weights of our discriminative model.

Experimental results We answer the following research questions:

- How often does our discriminative model for bilingual induction produce a correct translation within its top 10 guesses? Table 3 gives the top-10 accuracy for our model on Bengali, Tamil, Telugu, and Hindi, and shows its improvements over the standard unsupervised approach for combining multiple signals of translation equivalence.
- How much bilingual training data do we need in order to reach stable performance? We analyzed how accuracy changed as a function of the number

Language	MRR Baseline	Discriminative Model	Absolute Improvement	% Relative Improvement
Bengali	19.6	37.4	17.8	90.8
Tamil	17.1	37.9	20.8	121.6
Telugu	25.7	41.0	15.3	59.5
Hindi	25.9	43.4	17.5	67.6

Table 3: Top-10 Accuracy for bilingual lexicon induction on a test set. The accuracy increases significantly moving from the unsupervised MRR baseline to our discriminative model.

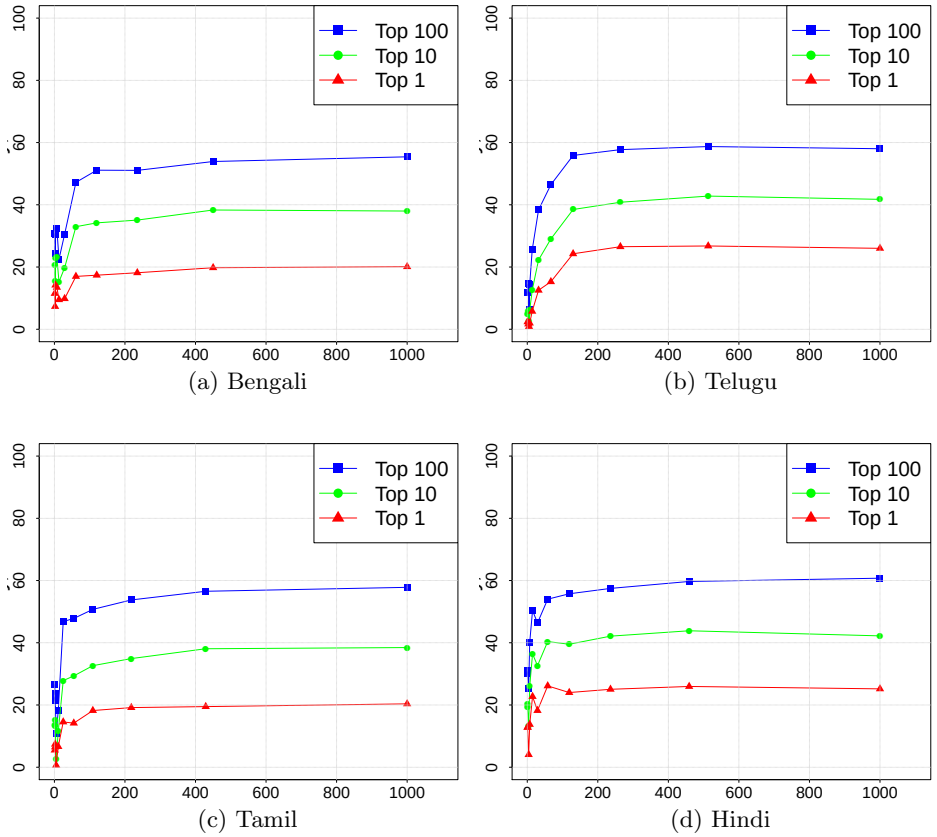
Source	গাণিতিকভাবে	ফাংশন	অভিষেক	পোষাকও	ফুটনোট	বোঝার
Induced Translations	mathematical equal ganitikovabe	**function** functions variables	made goal earned	shaky pashan shirts	mutant futbol futebol	vain newton boer
Correct Translation	mathematically	function	inauguration	dress	footnote	understand

Table 4: Examples of OOV Bengali words, our top-3 ranked induced translations, and their correct translations. Correct induced translations are bolded.

of bilingual dictionary entries used to train the discriminative model. Figure 4 shows learning curves that hold steady after approximately 300 training words.

- How much monolingual data would we need? Figure 5 shows a learning curve function of the size of the monolingual corpora used to estimate the similarity scores that are used as features in the model. The accuracy continues to increase, even beyond 10 million words. More monolingual data is better, but it is sometimes difficult to acquire even monolingual data in huge volumes for low resource languages.
- How well can our models translate rare words versus frequent words? Figure 6 shows that words that appear with higher frequency in our monolingual corpora tend to be translated better. (Pekar et al.2006) also investigated the effects of frequency on finding translations from comparable copper. This makes sense since we have more robust statistics when constructing their vector representations. The performance drops slightly for the highest frequency words, which are likely function words.

The effect of frequency has largely been ignored in past work on bilingual lexicon induction – most past work tried to discover translations only for the 1,000 most



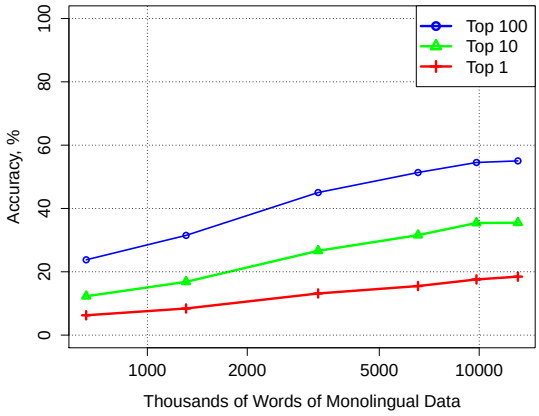
(a) Bengali (b) Telugu

(c) Tamil (d) Hindi

Fig. 4: Learning curves varying the number of dictionary entries used as positive training instances to our discriminative models, up to 1,000. For all languages, performance is fairly stable after about 300 positive training instances. The x-axis shows the number of dictionary entries used in training, and the y-axis gives the top-k accuracy of the model.

frequent words in a language.[4] The fact that low frequency words do not translate as well as high frequency words has significant implications for the application of bilingual lexicon induction to SMT. The most obvious use of learned translations would be as a way of augmenting what a SMT model learned from bitexts by applying bilingual lexicon induction to the OOV words. Unfortunately, the OOVs are lower frequency than the words that occurred in the bilingual training data. Therefore the translations are of mixed quality. Figure 4 shows some induced translations of Bengali words which were OOV with respect to a small bilingual training set.

[4] With some exceptions like (Pekar et al.2006) and (Daumé and Jagarlamudi2011), which tried to learn the translations of low-frequency words.



(a) Tamil

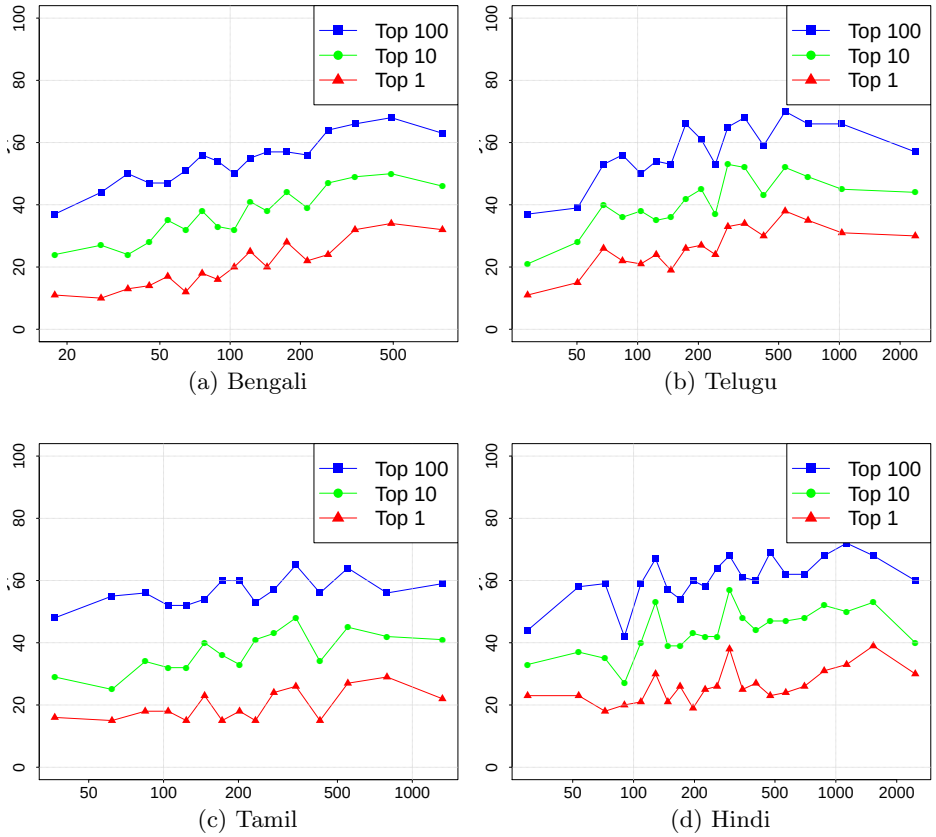
Fig. 5: Bilingual lexicon induction learning curves over varying monolingual corpora sizes for Tamil. The x-axis is shown on a log scale.

3 Transliterating OOV Words

Transliteration is a critical subtask of machine translation. Many named entities (NEs) (e.g. person names, organizations, locations) are transliterated rather than translated into other languages. That is, the sounds in the source language word are approximated with the target language phonology and orthography. Named entities constitute an open class of words. The names of people and organizations, for example, often show up in new documents and are often OOV with respect to the bilingual training data. Transliteration is therefore an alternative way of dealing with OOV items, and may produce more robust results than bilingual lexicon induction for NEs and cognates.

3.1 Our approach to transliteration

Following (Virga and Khudanpur2003), we treat transliteration as a monotone character translation task. Rather than using a noisy channel model, our transliteration models is based on the log-linear formulation of SMT described in (Och and Ney2002). Whereas SMT systems are trained on parallel sentences and use word-based n-gram language models, we use pairs of transliterated words along with character-based n-gram language models. We apply the word alignment algorithms from SMT to automatically align characters in pairs of transliterations. In fact, transliteration is simpler than translation, since phrases are often reordered in translation, but characters sequence are monotonic in transliteration. Our feature functions include a character sequence mapping probability (similar to the phrase translation probability), a character substitution probability (similar to the lexical probability), and a character-based language model probability. Table 5 shows some example transliteration rules that are learned using the SMT machinery.



(a) Bengali (b) Telugu

(c) Tamil (d) Hindi

Fig. 6: Bilingual lexicon induction accuracy as a function of source word **frequency** in Wikipedia monolingual data. Frequency is plotted along the x-axis. Top-k accuracy for the model is given in the y-axis.

Russian→English			
Rule	**Feature Function Scores**		
ф о м → f a u t	0.301	1.456	3.118
ц ы → t s y	0.204	2.490	1.431
щ у к → s c h u k	0.845	2.185	2.034
а р д ж с → a r j	0.398	1.432	0.506
Greek→English			
Rule	**Feature Function Scores**		
ο χ ΄α → o c h a	0.602	1.115	1.036
γ ε ρ → g e r	0.301	0.556	0.152
α λ μ → a l l m	0.699	0.214	0.175

Table 5: Examples of automatically learned transliteration rules from Russian to English and from Greek to English, along with their associated log probabilities for a character sequence mapping probability, a character substitution probability, and a character-based language model probability.

Bengali	2,100
Hindi	1,811
Malayalam	1,543
Tamil	1,463
Telugu	628
Urdu	893

Table 6: The number of Wikipedia articles with interlanguage links to English Wikipedia articles that describe people. These name pairs are used as training data to our SMT-inspired transliteration system.

3.2 Transliteration Experiments

Data We can use the standard SMT pipeline to learn transliteration rules, and we can produce transliterations of previously unseen words using an SMT decoder. The key is simply to find appropriate parallel data that shows transliterated pairs across different character sets (like between English's Roman alphabet and the Devanagari script used by Hindi). In (Irvine, Callison-Burch, and Klementiev2010), we detailed how we mined transliteration training data from Wikipedia page titles for 150 languages. Wikipedia's interlanguage links can be used as a source for example transliterations. We use the titles of non-Roman script languages that are paired with English pages that correspond to names. Wikipedia categorizes articles and maintains lists of all of the pages within each category. In mining transliteration data, we took advantage of a particular set of categories that list people born in a given year. For example, the Wikipedia category page '1961 births' includes links to the 'Barack Obama' and 'Michael J. Fox' pages. We iterated through birth years and the links to pages about people born in each year and then followed interlingual links from each English page about a person, compiling a large list of person names (Wikipedia page titles) in many languages. We found a total of 826,508 English Wikipedia pages about people. A similar process could be done to scrape other types of NEs, for instance by iterating over Wikipedia page categories for things like 'Countries in Africa' or 'Cities in Europe', but the expected yield would be lower than the number of person names. Table 6 gives the number of pairs of names between the English articles and the Indian languages that we examine in our end-to-end SMT experiments.

Experimental results Here we reproduce some of the experimental results from (Irvine, Callison-Burch, and Klementiev2010) that demonstrate the quality of our transliteration system. We evaluated our transliteration system on the ACL 2009 Named Entities Workshop, which featured a shared task on transliteration (Li et al.2009). The shared task evaluated systems trained to transliterate from English to several other languages using a variety of metrics. We used the workshop data to build a English-Hindi transliteration system, and compared our results against the other entries to the shared task. Table 7 shows our system's performance on the NEWS task – it is competitive with other systems entered into the shared task.

Metric	Our System	Other Systems
Top-1 Accuracy	.45	.00 – .50
Top-1 F-score	.87	.01 – .89
Mean Average Precision at 10	.18	.00 – .20

Table 7: A comparison of our performance (Irvine, Callison-Burch, and Klementiev2010) against the systems submitted to the Hindi transliteration shared task at the ACL 2009 Named Entities Workshop. There were 4,840 training pairs for English→Hindi in the NEWS shared task.

Candidate	Reference	Edit Dist	Normalized Edit Distance
Burkin	Burkin	0	.000
Andru**c**k	Andruk	1	.167
Shika**i**	S**c**hika**y**	2	.286
Gu**t**saev	Gu**z**a**y**ev	3	.427
Tr**u**xt**u**n	Tr**ak**st**o**n	4	.500

Table 8: Example transliterations. Sometimes the errors are near-misses where the system's proposed transliterations are only a few letters off from the reference transliteration. In these cases, the system does not receive any credit under metrics like the BLEU score, even though they may still be be useful for human readers. Normalized edit distance is the number of edits divided by the length of the reference.

Table 8 shows some example transliterations produced by our system paired with reference transliterations. Sometimes the system produces near-misses that could still be useful.

In our end-to-end translation experiments, we output the single best transliteration of each OOV word using our transliteration model. This transliteration was placed alongside the top-k translations proposed by the bilingual lexicon induction module. (Hermjakob, Knight, and Daumé III2008) trained a system so that it was able to learn when to transliterate versus translate. In our simpler setup, the SMT decoder had access to both transliterations and translations, and it used its model scores to select between the different options.

4 Building an End-to-End MT System with Small Parallel Corpora

The parameters of statistical models of translation are typically estimated from bilingual parallel corpora (Brown et al.1993). In (Klementiev et al.2012), we showed that it might be possible estimate the parameters of a phrase-based SMT system from monolingual corpora instead of a bilingual parallel corpus. We replaced the standard features from the phrase-based models (such as the phrase translation

probabilities) with the monolingual signals of translation equivalence used in bilingual lexicon induction (Section 2). In the (Klementiev et al.2012) study, we worked with estimating the parameters from Spanish-English, and we had an idealized scenario in that we performed bilingual lexicon induction on two halves of a bilingual parallel corpus. We further showed that keeping all of the standard bilingually estimated features and adding monolingually estimated features from bilingual lexicon induction seemed to improve the translation quality over bilingual features alone.

In this section, we do a deeper analysis of the experiments that we originally published in (Irvine and Callison-Burch2013a). We enhanced the phrase tables for 6 low-resource Indian languages (translating Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu into English). We examine two ways of improving the the quality of low-resource machine translation:

- We add translations of OOV words (and of low-frequency words) using our discriminative bilingual lexicon induction model. This allows better coverage by the models of the words in the test set that do not appear, or appear only rarely, in the training data.
- We incorporate new features into the SMT model based on the different signals of translation equivalence that we use our bilingual lexicon induction method. The features are included both for monolingually induced translations, and for translations learned from the small bitexts. The features are combined in a log linear model, and their weights are set using batch MIRA (Cherry and Foster2012).

For all 6 languages, we see improvements in translation quality, ranging from 0.6 and 1.7 BLEU points. These experiments represent a realistic way of improving SMT using bilingual lexicon induction for genuinely low resource languages.

4.1 Data

(Post, Callison-Burch, and Osborne2012) used MTurk to collect small parallel corpora for the following Indian languages and English: Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu. They collected both parallel sentence pairs and a dictionary of word translations. We use all six datasets, which provide real low resource data conditions for six truly low resource language pairs. Tables 9 and 10 show statistics about the datasets.

As usual, we use both our web crawls and our Wikipedia comparable corpora for each language pair. Dataset sizes are given in Table 2 for Bengali, Hindi, Tamil and Telugu. For Malayalam, we had 4 million words in our web crawled data, and 5 million words in our Wikipedia data (with 17,000 interlanguage links). For Urdu, we had 285 million words in our web crawled data, and 3 million words in our Wikipedia data (with 15,000 interlanguage links).

4.2 Experimental setup

We use the training/development/test data splits given by (Post, Callison-Burch, and Osborne2012) and, following that work, include the dictionaries in the training

Language	Words of Training Data (from Sentences)	(from Dictionary)	Dev Types % OOV	Dev Tokens % OOV
Tamil	334,714	77,240	44	25
Telugu	414,094	40,742	39	21
Bengali	239,555	6,783	37	18
Malayalam	263,086	151,194	6	3
Hindi	658,977	0	34	11
Urdu	615,635	116,496	23	6

Table 9: Information about datasets released by (Post, Callison-Burch, and Osborne2012): words in the source language parallel sentences and dictionaries, and percent of development set word types and tokens that are OOV (do not appear in either section of the training data). (Post, Callison-Burch, and Osborne2012) did not provide a dictionary for Hindi, so we exclude it from the baseline SMT system.

data and report results on the devtest set using case-insensitive BLEU and four references. We use the Moses phrase-based MT framework (Koehn et al.2007). For each language, we extract a phrase table with a phrase limit of seven. In order to make our results comparable to those presented in (Post, Callison-Burch, and Osborne2012), we follow that work and use the English side of the training data to train a language model. Using a language model trained on a larger corpus (e.g. the English side of our comparable corpora) may yield better results, but such an improvement is orthogonal to the focus of this work. Throughout our experiments, we use the batch version of MIRA (Cherry and Foster2012) for tuning the feature set. We rerun tuning for all experimental conditions and report results averaged over three tuning runs (Clark et al.2011).

Our baseline uses the bilingually extracted phrase pairs and standard translation probability features. We augment it with the single top ranked translation for each OOV to improve coverage (+ OOV Trans) and with additional features to improve accuracy (+Features). We make each modification separately and then together. Then we present additional experiments where we induce translations for low frequency words, in addition to OOVs (4.2.2), append top-k translations (4.2.3), vary the amount of training data used to induce the baseline model (4.2.4), and vary the amount of comparable corpora used to estimate features and induce translations (4.2.5).

Results: Bilingual Lexicon Induction Before presenting end-to-end MT results, we examine the performance of the supervised bilingual lexicon induction technique that we use for translating OOVs. In Table 11, top-1 accuracy is the percent of source language words in a held out portion of the training data[5] for which the highest ranked English candidate is a correct translation. (Post, Callison-Burch,

[5] We retrain with all training data for MT experiments.

Language Pair	Training	Development	Test
Bengali-English	20,788	914	1,001
Hindi-English	37,726	1,082	1,113
Malayalam-English	29,518	1,166	1,267
Tamil-English	35,027	1,292	1,225
Telugu-English	43,038	1,263	1,047
Urdu-English	33,798	736	605

Table 10: The number of sentence pairs in the training/dev/test set splits for the Indian-language bilingual parallel corpora released by (Post, Callison-Burch, and Osborne2012).

Language	Top-1 Acc.	Top-10 Acc.
Tamil	4.5	10.2
Telugu	32.8	47.9
Bengali	17.9	29.8
Malayalam	12.9	23.0
Hindi	44.3	57.6
Urdu	16.1	33.8

Table 11: Percent of word types in a held out portion of the training data which are translated correctly by our bilingual lexicon induction technique. Evaluation is over the top-1 and top-10 outputs in the ranked lists for each source word.

and Osborne2012) gathered up to six translations for each source word, so some have multiple correct translations. Performance is lowest for Tamil and highest for Hindi. For all languages, top-10 accuracy is much higher than the top-1 accuracy. In Section 4.2.3, we explore appending the top-k translations for OOV words to our model instead of just the top-1.

4.2.1 Improving Coverage and Accuracy in End-to-End SMT

Table 12 shows our results adding OOV translations, adding features, and then both. Simply adding monolingually estimated features functions to the phrase table improves our models' accuracy, increasing BLEU scores between 0.18 (Bengali) and 0.60 (Malayalam).

Adding OOV translations makes a big difference for some languages, such as Bengali and Urdu, and almost no difference for others, like Malayalam and Tamil. The OOV rate (Table 9) is low in the Malayalam dataset and high in the Tamil dataset. However, as Table 11 shows, the translation induction accuracy is low for both. Since few of the supplemental translations are correct, we don't observe BLEU gains. In contrast, induction accuracies for the other languages are higher,

	Baseline	+Features	+OOV Trans.	+Features & Trans
Tamil	9.5	9.8	9.5	**10.0**
Telugu	11.7	12.0	12.2	**12.3**
Bengali	12.1	12.3	**12.7**	12.6
Malayalam	13.6	14.2	13.7	**14.2**
Hindi	15.0	15.3	15.6	**16.1**
Urdu	20.4	21.0	21.3	**21.8**

Table 12: BLEU scores improve for all 6 low resource languages when we add translations of OOV using bilingual lexicon induction (+OOV Trans.), and when we add monolingually-derived features to the standard phrase table features (+Features). The greatest gains come from incorporating both OOV translations and new features (+Features & Trans).

OOV rates are substantial, and we do observe moderate BLEU improvements by supplementing phrase tables with OOV translations.

Combining the two methods results in translations that are better than applying either technique alone for five of the six languages. BLEU gains range from 0.5 (Bengali) to 1.4 (Urdu). We attribute the particularly good Urdu performance to the relatively large monolingual corpora (Table 2). In Section 4.2.5, we present results varying the amount of Urdu-English comparable corpora used to induce translations and estimate additional features.

4.2.2 Translations of Low Frequency Words

Beyond adding translations just for strictly OOV words, we wanted to evaluate whether bilingual lexicon induction could also be useful for low frequency words. Strictly speaking, adding translations of OOV words will never decrease the BLEU score, since even adding in a random translation is no worse (under BLEU) than outputting a foreign word written in a non-Roman script.

For source words which only appear a few times in the parallel training text, the bilingually extracted translations in the standard phrase table are likely to be inaccurate and incomplete. Augmenting a model with additional translations for low frequency words may fix some other types of errors, for instance a source word was observed in training with a translation that is not the correct sense for the test set.

We perform additional experiments varying the minimum source word training data frequency for which we induce additional translations. That is, if $freq(w_{src}) \leq M$, we induce a new translation for it and include that translation in our phrase table. Note that in the results presented in Table 12, $M = 0$, meaning that it only adds induced translations for OOVs and not for low frequency words that occurred once or more in the training data. In these experiments, we include our

Language	Baseline	M: trans added for $freq(w_{src}) \leq M$					
		0	1	5	10	25	50
Tamil	9.5	10.0	9.9	**10.2**	**10.2**	9.9	**10.2**
Telugu	11.7	12.3	12.2	12.3	**12.4**	12.3	11.9
Bengali	12.1	12.6	12.8	13.0	12.9	**13.1**	13.0
Malayalam	13.6	**14.2**	14.1	**14.2**	**14.2**	13.9	13.9
Hindi	15.0	16.1	16.1	**16.2**	**16.2**	16.0	15.8
Urdu	20.4	21.8	21.8	21.8	21.9	**22.1**	21.8

Table 13: Varying minimum parallel training data frequency of source words for which new translations are induced and included in the phrase-based model. In all cases, the top-1 induced translation is added to the phrase table and features estimated over comparable corpora are included (i.e. +Feats & Trans model).

additional phrase table features estimated over comparable corpora and hope that these scores will assist the model in choosing among multiple translation options for low frequency words, one or more of which is extracted bilingually and one of which is induced using comparable corpora. Table 13 shows the results when we vary M. As before, we average BLEU scores over three tuning runs.

In general, modest BLEU score gains are made as we augment our phrase-based models with induced translations of low frequency words. The highest performance is achieved when M is between 5 and 50, depending on language. The largest gains are 0.5 and 0.3 BLEU points for Bengali and Urdu, respectively, at $M = 25$. This is not surprising; we also saw the largest relative gains for those two languages when we added OOV translations to our baseline model. With the addition of low frequency translations, our highest performing Urdu model achieves a BLEU score that is 1.7 points higher than the baseline.

In different data conditions, inducing translations for low frequency words may result in better or worse performance. For example, the size of the training set impacts the quality of automatic word alignments, which in turn impacts the reliability of translations of low frequency words. However, the experiments detailed here suggest that including induced translations of low frequency words will not hurt performance and may improve it.

4.2.3 Appending Top-K Translations

So far we have only added the top-1 induced translation for OOV and low frequency source words to our phrase-based model. However, the bilingual lexicon induction results in Table 11 show that accuracies in the top-10 ranked translations are, on average, nearly twice the top-1 accuracies. Here, we explore adding the top-k induced translations. We hope that our additional phrase table features estimated over comparable corpora will enable the decoder to correctly choose between the

Language	Baseline	*k*: top-*k* translations added				
	1	3	5	10	25	
Tamil	9.5	**10.0**	**10.0**	9.8	**10.0**	**10.0**
Telugu	11.7	**12.3**	11.7	11.9	11.7	11.6
Bengali	12.1	12.6	12.6	12.6	12.7	**12.8**
Malayalam	13.6	**14.2**	**14.2**	**14.2**	**14.2**	14.1
Hindi	15.0	**16.1**	16.0	15.9	15.9	15.9
Urdu	20.4	**21.8**	**21.8**	21.7	21.5	21.6

Table 14: Adding top-k induced translations for source language OOV words, varying *k*. Features estimated over comparable corpora are included (i.e. +Feats & Trans model). The highest BLEU score for each language is highlighted. In many cases differences are less than 0.1 BLEU.

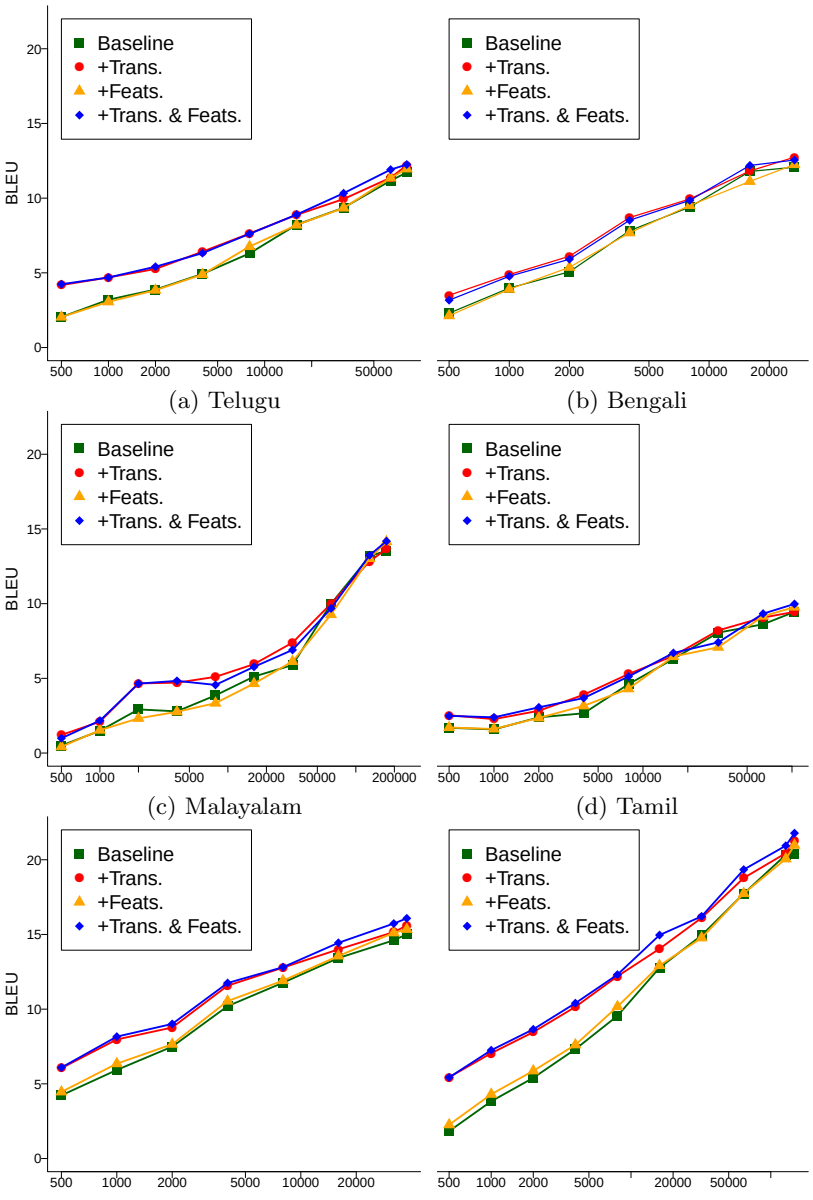
k translation options. We induce translations for OOV words only ($M = 0$) and include all comparable corpora features.

Table 14 shows performance as we append the top-k ranked translations for each OOV word and vary *k*. With the exception of Bengali, using a k greater than 1 does not increase performance. In the case of Bengali, and additional 0.2 BLEU is observed when the top-25 translations are appended. In contrast, we see performance decrease substantially for other languages (0.7 BLEU for Telugu and 0.2 for Urdu) when the top-25 translations are used. Therefore, we conclude that, in general, the models do not sufficiently distinguish good from bad translations when we append more than just the top-1. Although using a k greater than 1 means that more correct translations are in the phrase table, it also increases the number of possible outputs over which the decoder must search.

4.2.4 Learning Curves over Parallel Data

In the experiments above, we only evaluated our methods for improving the accuracy and coverage of models trained on small amounts of bitext using the full parallel training corpora released by (Post, Callison-Burch, and Osborne2012). Here, we apply the same techniques but vary the amount of parallel data in order to generate learning curves. Figure 7 shows learning cures for all six languages. In all cases, results are averaged over three tuning runs. We sample both parallel sentences and dictionary entries.

All six learning curves show similar trends. In all experimental conditions, BLEU performance increases approximately linearly with the log of the amount of training data. Additionally, supplementing the baseline with OOV translations improves performance more than supplementing the baseline with additional phrase table scores based on comparable corpora. However, in most cases, supplementing the baseline with both translations and features improves performance more than either alone. Performance gains are greatest when very little training data is used. The Urdu



(a) Telugu (b) Bengali

(c) Malayalam (d) Tamil

(e) Hindi (f) Urdu

Fig. 7: Comparison of learning curves over lines of parallel training data for four SMT systems: our baseline phrase-based model (baseline), model that supplements the baseline with translations of OOV words induced using our supervised bilingual lexicon induction framework (+Trans), model that supplements the baseline with additional phrase table features estimated over comparable corpora (+Feats), and a system that supplements the baseline with both OOV translations and additional features (+Trans & Feats).

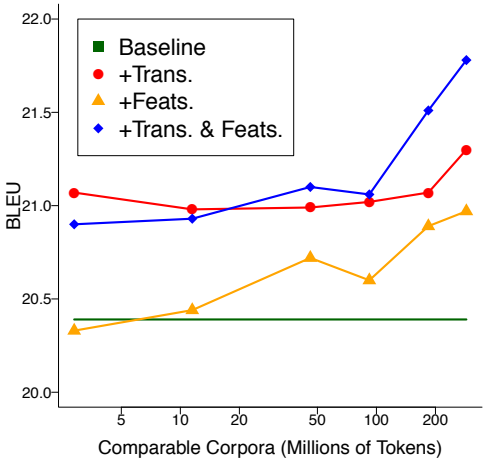


Fig. 8: Urdu to English translation results using varying amounts of monolingual corpora to estimate features and induce translations.

learning curve shows the most gains as well as the cleanest trends across training data amounts. As before, we attribute this to the relatively large comparable corpora available for Urdu.

4.2.5 Learning Curves over Comparable Corpora

In our final experiment, we consider the effect of the amount of *bilingual comparable corpora* that we use to estimate features and induce translations. We present learning curves for Urdu-English because we have the largest amount of monolingual corpora for that pair. We use the full amount of parallel data to train a baseline model, and then we randomly sample varying amounts of our Urdu-English monolingual corpora. Sampling is done separately for the web crawl and Wikipedia comparable corpora. Figure 8 shows the results. As before, results are averaged over three tuning runs.

The phrase table features estimated over comparable corpora improve end-to-end MT performance more with increasing amounts of comparable corpora. In contrast, the amount of comparable corpora used to induce OOV translations does not impact the performance of the resulting MT system as much. The difference may be due to the fact that data sparsity is always more of an issue when estimating features over *phrase pairs* than when estimating features over *word pairs* because phrases appear less frequently than words in monolingual corpora. Our comparable corpora features are estimated over phrase pairs while translations are only induced for OOV words, not phrases. So, it makes sense that the former would benefit more from larger monolingual corpora.

5 Building an End-to-End MT System with Zero versus Small Bitexts

In this section, we build several end-to-end Hindi-English SMT systems. We use a variety of techniques to construct the translation models, including using existing (incomplete) bilingual dictionaries to gloss the text, and using our transliteration model and our bilingual induction model to translate OOV words. We translate two (arbitrary chosen) Hindi Wikipedia pages about Islam and Forests. We qualitatively evaluate the output our a system that uses no bilingual sentence-aligned parallel corpora and compare it to the output of models trained with small amounts of bitexts.

We generate a Hindi-English phrase tables in the following way:

- We add all entries from the Hindi-English bilingual dictionaries. The existence of a bilingual dictionary is more likely than a large sentence-aligned bilingual parallel corpus, and it is required for our bilingual lexicon induction model.
- We generate the 1-best transliteration for all non-Roman script words in the Hindi articles. As we described in Section 3, we do transliteration by training character-based translation models on Wikipedia page titles.
- We generate the top-10 translations for all OOV Hindi words using our bilingual induction model.
- Then, we score each patchwork phrase table using the following similarity features: web crawl contextual similarity, web crawl temporal similarity, Wikipedia contextual similarity, Wikipedia topic similarity, and orthographic similarity.

In addition to generating the phrase table, we also use a language model computed over the entire English Wikipedia, except for the English versions of the pages which we wish to translate.

Typically in SMT, in addition to using parallel corpora to estimate the parameters of an SMT model, a small bitext is also used as a development set to tune the feature weights of the log linear model. Since we are assuming a zero bitext setting here, we also assume that there is no such data available for tuning. Rather than tuning the parameters specifically for Hindi-to-English, we reuse the weights that were learned for a Bengali-to-English English MT experiment that used the same set of monolingually derived features. The choice to re-use the model parameters from Bengali rather than some other language was arbitrary. Of course, the source language and corpora change substantially in these new experiments, and the optimal weights are unlikely to be the same.

Rather than evaluating these translations with an automatic metric like BLEU, we show example translations and evaluate them qualitatively. Because the topics are familiar, it is possible to read the output and get a sense of the translation quality. Figures 9 and 10 (pages 30 and 31) show the first few sentences of the Hindi Wikipedia pages on *Forest* and *Islam* translated several ways. In each figure, we show the Hindi source paragraph and the different ways that it can be rendered into English. The figures show:

(1) The original Hindi paragraph.

(2) The dictionary to gloss the Hindi words into English. The dictionary gloss is based on bilingual dictionaries that we collected on MTurk. If the dictionary contain more than one translation of a given word, we pick one randomly. The dictionary glosses are somewhat readable, but there are many OOV words.

(3) A transliteration of each of the Hindi words into Roman script. Although the transliterations of some cognates, including *hayadrologik* and *biosphia* in the forest translation, are understandable, most words are not. The number of cognates and named entities, which can often be accurately transliterated instead of translated, varies by subject matter. For example, in the Hindi page on Barack Obama, there are many more 'transliteratable' words than the Hindi page on forests.

(4) Here we construct a phrase table with the dictionary translations and transliterations. The monolingually-derived scores allow us to select between dictionary translations and transliterations (and to select between alternate translations when there are multiple entries in the bilingual dictionary). The results are much better translations than either gloss. For translations 4-7, we use a 5-gram language model trained on the English gigaword corpus.

(5) Here we construct the phrase table not only with the dictionary translations and transliterations, but also with the top-10 translations that we induce for each Hindi word by the bilingual induction model presented in Section 2. This is a full transliteration model estimated using no parallel training data whatsoever. Introducing induced translations has several noticeably positive effects. For example, in the first sentence of the forest translation, the transliterations *uchucha*, *esjangal*, and *podahe* are used in the 'Dictionary + Transliterations + Monolingual Scoring' model. Here we instead use the corresponding induced translations *systolic*, *canopy*, and *headless* instead of the non-sensical transliterations. None of these words is a completely accurate translation, but they are closer than the non-sensical transliterations. This condition represents the most complete system that we can build with zero bitexts.

(6) This translation is produced by the model trained on our small Hindi training bitext (used in Section 4). This is the type of translation that results from running standard SMT on low resource language pairs. There is a relatively high OOV rate, but the words that are seen in the bitext are translated fairly reasonably.

(7) The final translation takes advantage of our entire bag of tricks: the small training bitext, our bilingual dictionaries, transliterations, induced translations, and monolingual scoring. The phrase table is populated with the top-10 induced translations, top-1 transliterations, dictionary pairs, and phrase pairs extracted from the word aligned training text. Each phrase pair is scored monolingually and those taken from the bitext are also scored bilingually. Like the dictionary word gloss, using the model trained on the small bitext to translate the Hindi text alone results in many OOV words. However, using the small bitext allows us to accurately translate function words plus common words and phrase, for example *which is* and *one of the most important.*

(8) A human reference translation.

Qualitatively, we prefer the final automatic translation (7) over the other automatic translations (2-6) for the Hindi articles about forests and Islam. This model takes advantage of both bilingual and monolingual resources. Although the translations are certainly not publishable in any of the conditions, they are useful for understanding the gist of the text, and even the zero-bitext translation (5) might be useful for downstream NLP technologies like topic detection and tracking systems (Church and Hovy1993).

6 Discussion and Other Related Work

In this article, we have assumed that a bilingual dictionary is available. Several past efforts have tried to eliminate even this assumption. Notably, (Ravi and Knight2011) built a full machine translation system using decipherment techniques. With these techniques, they are able to produce translations without bilingual parallel corpora, and without bilingual dictionaries.

Other approaches to bilingual lexicon induction attempt to do away with the requirement of having a seed bilingual dictionary. (Vulić, De Smet, and Moens2011) propose a bilingual Latent Dirichlet Allocation model for finding translations from comparable corpora without using any other linguistic resources. Other bilingual lexicon induction techniques, from (Koehn and Knight2002) and (Peirsman and Padó2010), have tried to solve the problem of projecting across the vector spaces for two languages by seeding with orthographically similar words instead of small bilingual dictionaries. (Vulić and Moens2013) presents a systematic study of different ways of bootstrapping the projection across the vector spaces of two languages. (Chu, Nakazawa, and Kurohashi2014) also does away with the seed bilingual dictionary by first using topic models to find similar words, and then using those as the seed to a context-based model.

7 Conclusion

In this article, we used bilingual lexicon induction techniques to create and rescore phrase tables for a machine translation system for low resource languages. We pushed the idea of learning translations from monolingual corpora to its logical conclusion by building a full end-to-end machine translation system without any of the sentence-aligned bilingual parallel training data that is typically required by SMT systems. We additionally demonstrated that the induced translations and the associated scoring techniques can be used to improve the quality of SMT when we have only small amounts of parallel text to train our translation models. Rather than simulating a low-resource setting, we undertook the task of translating truly low resource languages.

8 Acknowledgments

This material is based on research sponsored by DARPA under contract HR0011-09-1-0044 and by the Johns Hopkins University Human Language Technology Center

of Excellence. 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.

We would like to thank David Yarowsky for his tremendous support, and for his inspiring work on –and continued ideas about– learning translations from monolingual texts. We would like to thank Alex Klementiev for his substantial contributions to this research and his comments on a draft of this article. We would like to thank Manaal Faruqui and Sneha Jha for providing the reference translations for the two Hindi paragraphs. Thank you to the two anonymous reviewers who provided valuable feedback on the first draft of this manuscript.

Original Text (1)

एक वन एक उच्च घनत्व के साथ एक क्षेत्र है पेड़ (tree) एसजंगल के कई परिभाषाएँ , है जो की विभिन्न मानदंडों पर आधारित हैं.यह पोदाहे लगभग ९.४ % पृथ्वी की सतह को घेर ते है (या 30 %) जो की आवासों (habitat) ह्यद्रोलोगिक प्रवाह (hydrologic flow) मोदुलातोर्स (modulator) , और मिट्टी (soil) बचाव , एक पृथ्वी के बीओस्फिअ का सर्वाधिक महत्वपूर्ण पहलुओं के गठन.का प्रवास करते है इतिहास बताता है , की " वन " एक बीहड़ क्षेत्र जिसका मतलब कानूनी तौर पर बाजू के लिए निर्धारित शिकार (hunting) के द्वारा सामंती (feudal) कुलीनता (nobility) है , और इन शिकार जंगलों जरूरी ज्यादा अगर में सभी (देखें जंगली नहीं थे रॉयल वन (royal forest)) .हालांकि , शिकार के जंगलों अक्सर वुडलैंड के महत्वपूर्ण क्षेत्रों को शामिल किया जबकि , शब्द वन अंततः जंगली भूमि अधिक सामान्यतः मतलब करने के लिए आया था.एक वुडलैंड (woodland) जो की एक जंगल से भिन्न है .

Dictionary Word Glosses (2)

one forest one उच्च density its साथ one field is wood (tree) एसजंगल its lots definitions , is which of various मानदंडों on based ह.यह पोदाहे total ९.४ % the earth of surface को surround ते is (either 30 %) those of आवासों (habitat) ह्यद्रोलोगिक प्रवाह (hydrologic flow) मोदुलातोर्स (modulator) , and soil (soil) safeguard , one the earth its बीओस्फिअ का rules important sides of गठन.का foreign do is history telling is , of " forest " one बीहड़ field whose means कानूनी for on बाजू of for nidhirit शिकार (hunting) its द्वारा सामंती (feudal) कुलीनता (nobility) is , and these शिकार in jungles compulsory more if me all (see wild no was royal forest (royal forest)) .हालांकि , शिकार its in jungles usual वुडलैंड its importance areas को शामिल did while , शब्द forest at the end wild land more generally means do of for was था.एक वुडलैंड (woodland) which of one जंगल from different is .

Transliteration Gloss (3)

ak vn ak uchcha ghantwa ke sath ak ksatra ha ped (tree) esjangal ke ki pribhashaën , ha jo ki vibhinn mandndon pr adharit han.yh podahe lgbhag . % prithvi ki sath ko gher te ha (ya 30 %) jo ki avason (habitat) hayadrologik prawah (hydrologic flow) modulators (modulator) , mitti (soil) bchaw , ak prithvi ke biosphia ka sarveadhik mahatwpurn phluon ke gthn.ka prawas krate ha dharampal battata ha , ki " vn " ak bihd ksatra jiska mtalb kanuni taur pr baju ke lier nirdharit shikar (hunting) ke dwara samanti (feudal) kulenta (nobility) ha , in shikar junglon jruri jayada agar man sbhi (dekhen jungali nhin the royle vn (royal forest)) .hallanki , shikar ke junglon aksr woodland ke mahatwpurn ksatron ko shamill kiya jbki , shbd vn antt: jungali bhumi adhik samanyat: mtalb karne ke lier aya tho.aq woodland (woodland) jo ki ak jangal se bhinn ha .

Dictionary + Transliterations + Monolingual Scoring (4)

one forest one uchcha density of sath one field is tree (tree) esjangal of many definitions , is which of various mandndon on based han.yh podahe nearly . % of the earth surface ko surround te is (or 30 % of which) avason (habitat) hayadrologik prawah (hydrologic flow) modulators (modulator) , and soil (soil) safeguard , one of the earth biosphia ka more important sides of gthn.ka foreign to do is history telling is , of " forest " one bihd field whose means kanuni for on its baju for nidhirit shikar (hunting) of dwara samanti (feudal) kulenta (nobility) is , and these shikar forests necessary more if among all (see no wild was royal forest (royal forest)) .hallanki , shikar of forests often woodland of important areas ko shamill did while , shbd forest at the end wild land more generally means do its for was tho.aq woodland (woodland) which of one jangal from different is .

Dictionary + Transliterations + Induced + Monolingual Scoring (5)

one forest one systolic density of which one field is tree (tree) canopy of many definitions , is which of various crm on based han.yh nearly headless . % of the earth surface ko surround te is (or 30 %) which of keyhole (organisms) canopy irr (telecom low) modulators (coniferous) , and soil (erosion) safeguard , one the earth of app ka more important sides of gthn.ka foreign to do is history telling is , of " forest " one maestra field whose means responsibility for on pulleys of for nidhirit mane (africana) of dhara necker (electors) émigrés (forest) is , and these lions forests more necessary if among all (see no wild the royal forest (royal society)) .hallanki , mane of forests often evergreen of important areas ko they did while , quirk forest at the end wild land more generally means do its for was tho.aq evergreen (forests) which of one forest from different is .

Small Bitext Translation (6)

a forest with a high density is one area of the tree (एसजंगल) many definitions of the tree , which is full of various मानदंडों पोदाहे based on almost ९.४ % to the surface of the earth is ' घेर) (or 30 % of the habitat , which produced (flow)) (flow ह्यद्रोलोगिक (hydrologic) मोदुलातोर्स (modulator) soil , and the will of the earth , one of the most important scripts बीओस्फिअ गठन.को . history tells everybody , " " the " " forest , which a बीहड़ area on legal means for बाजू (hunting victim decided by the feudal) कुलीनता ((सामंती)) nobility , and these victim is more important जंगलों (, forests , see if all in were not royal forest .हालांकि ( )) royal forest of the victim , often वुडलैंड जंगलों included to the important areas , while in the forest , word means more generally , forest land for वुडलैंड came to woodland ( ) था.एक , which is different from the one from .

Small Bitext Trans + Dict + Translit + Induced + Mono Scoring (7)

a high density of a forest area is with a tree (tree) canopy of definitions , which is one of the many different crm based on this canopy almost . % of the earth 's surface to surround ' is (or 30 %) which houses , (elephants) canopy flow (hydrologic low) canopy (canopy) , and the soil (erosion) saved , one of the most important earth 's monoculture gthn.ka journey of scripts . " " he tells the history , " " a forest area maestra , whose legal means for the pulley on decided victim (elephants) by tong (electors) danton (forest) , and more important forests victim if these all in (see was not , wild royal forest (royal forest)) .hallanki , victim of the evergreen often forests of important areas , while . finally , wild forest land , the word means more commonly used to refer to came tho.aq evergreen (forests) , which is different from the one from .

Reference Human Translation (8)

A forest is an area with a high density of trees. There are many definitions of a forest, which depend on various standards. These plants cover approximately 9.4% of the area of earth or 30% of habitats, hydrologic flow modulators, soil protection, constitute the most important aspects of the Earth's biosphere. History tells us that a forest is a wilderness area which means legally designated for hunting by feudal nobility. In these royal hunting forest were not wild forests. Though, hunting forests have often been included in important areas of woodland still, the word forest finally was generally used to refer to wild land. A woodland is different from a forest.

Fig. 9: First paragraph of Hindi Wikipedia page on *Forest*, and a progression of translations of it.

Original Text (1)

इस्लाम धर्म (الإسلام।) ईसाई धर्म के बाद अनुयाइयों के आधार पर दुनिया का दूसरा सब से बड़ा धर्म है । इस्लाम शब्द अरबी भाषा का शब्द है जिसका मूल शब्द सल्लमा है जिस की दो परिभाषाएं हैं (१) अमन और शांति (२) आत्मसमर्पण । ईस्लाम एकेश्वरवाद को मानता है । इसके अनुयायियों का प्रमुख विश्वास है कि ईश्वर सिर्फ़ एक है और पूरी सृष्टि में सिर्फ़ वह ही महिमा (इबादत) के लायक है , और सृष्टि में हर चीज , ज़िंदा और बेजान , दृश्य और अदृश्य उसकी इच्छा के सामने आत्मसमर्पित और शांत है । इस्लाम धर्म की पवित्र पुस्तक का नाम कुरआन है जिसका हिंदी में मतलब सस्वर पाठ है । इसके अनुयायियों को अरबी में मुस्लिम कहा जाता है , जिसका बहुवचन मुसलमान होता है । मुसलमान यह विश्वास रखते है कि कुरआन जिब्राईल (ईसाईयत में gabriel) नामक एक फ़रिश्ते के द्वारा , मुहम्मद साहब को ७वीं सदी के अरब में , लगभग २३ साल में याद-कंठस् थ कराया गया था । मुसलमान इस्लाम को कोई नया धर्म नहीं मानते । उनके अनुसार ईश्वर ने मुहम्मद साहब से पहले भी धरती पर कई दूत भेजे हैं , जिनमें इब्राहीम , मूसा और ईसा शामिल हैं । मुसलमानों के अनुसार मूसा और ईसा के कई उपदेशों को लोगों ने विकृत कर दिया । अधिकतम मुसलमानों के लिये मुहम्मद साहब ईश्वर के अन्तिम दूत थे और कुरआन मनुष्य जाति के लिये अन्तिम संदेश है ।

Dictionary Word Glosses (2)

islam religion (الإسلام।) christian religion of after अनुयाइयों its foundation / support on world का another all from huge religion is । islam शब्द arabic language का शब्द is whose worth शब्द सल्लमा is from which of दो परिभाषाएं are (१) अमन and peace (२) dedicated ।
ईस्लाम एकेश्वरवाद को believe is । for this followers का major विश्वास is that god सिर्फ़ one is and complete universe me सिर्फ़ that only महिमा (worship) its लायक is , and universe me every thing , ज़िंदा and बेजान , view and invisible his desire its front आत्मसमर्पित and quiet is । islam religion of holy book का name kuran -holy book of islam is whose hindi me means सस्वर lesson is । for this followers को arabic me मुस्लिम कहा go is , whose बहुवचन muslim happens is । muslim यह विश्वास रखते is that koran जिब्राईल (ईसाईयत among gabriel) नामक one फ़रिश्ते of द्वारा , mohammad saheb को ७वीं century its arab me , total 23 year me याद-कंठस् tha कराया went था । muslim islam को कोई नया religion no know as । his अनुसार god ने mohammad saheb from before also earth on lots embassdor sent are , in which इब्राहीम , musa and ईसा शामिल is । muslims its अनुसार musa and ईसा of many lectures को people ने विकृत do gave । maximum muslims its for mohammad saheb god of अन्तिम embassdor was and koran मनुष्य race of for अन्तिम message is ।

Transliteration Gloss (3)

islam dharam (الإسلام।) isai dharam ke bad anuyaiyon ke adhar pr dunia ka dusara sb se bdha dharam ha . islam shbd arbi bhasha ka shbd ha jiska moole shbd sallama ha gis ki do pribhashaën han (i) aman shanti (ii) atamsamarpn .
islam akeshwarvad ko manta ha . iske anuyayion ka pramukh viswaas ha ki ishwar sierf ak ha puri s.ti man sierf vh hi mahima (ibadt) ke laik ha , s.ti man har chez , zinda began , drishy adrishy usky iachha ke samane atamsamarpit shant ha . islam dharam ki pavitra pustek ka nam quran ha jiska hindi man mtalb sswar path ha . iske anuyayion ko arbi man muslim kha jata ha , jiska bahuvchn musalman hota ha . musalman yh viswaas rkhte ha ki quran gibrail (isaiyat man gabriel) namk ak frishte ke dwara , muhammad sahb ko viiwin sdi ke arb man , lgbhag ii sal man yad-kanthus th karaya ghiya tha . musalman islam ko koi nya dharam nhin manate . unce anussaur ishwar ne muhammad sahb se phle bhi dhrati pr ki dut bhage han , jinman ibrahim , musa isa shamill han . musalmanon ke anussaur musa isa ke ki upadeshon ko logon ne vicrit kar dia . adhictam musalmanon ke lier muhammad sahb ishwar ke anthim dut the quran manushy jati ke lier anthim sandesh ha .

Dictionary + Transliterations + Monolingual Scoring (4)

islam religion (الإسلام।) christian religion of after anuyaiyon its basis on world ka second all from big religion is . islam shbd arabic language ka shbd is whose original shbd sallama is from which of do pribhashaën is (1) aman and peace (ii) dedicated .
islam akeshwarvad ko believe is . its followers ka major viswaas is that god sierf one is and complete universe among sierf that only mahima (worship) of laik is , and universe among each and every thing zinda , and began , view and invisible his desire of front atamsamarpit and quiet is . islam religion of holy book ka name quran is whose hindi among means sswar path is . its followers ko arabic among muslim kha go is , whose bahuvchn muslim happens is . muslim yh viswaas rkhte is that koran gibrail (isaiyat among gabriel) namk one frishte of dwara , muhammad saheb ko viiwin century of arab man , nearly 23 year among yad-kanthus tha karaya made tha . muslim islam ko koi nya no religion feel . his anussaur god ne muhammad saheb from before also earth on many dut is sent , in which ibrahim , musa and isa shamill is . muslims of anussaur musa and isa of many lectures ko people ne vicrit do give . maximum muslims of for muhammad saheb god of anthim dut and the koran manushy race for its anthim message is .

Dictionary + Transliterations + Induced + Monolingual Scoring (5)

islam religion (alevi) christian religion of after adulation of basis on world ka second all from big religion is . islam quirk arabic language ka quirk is whose original quirk isis is from which of do rima is (1) aman and peace (ii) dedicated .
islam anic ko believe is . its followers ka major undead is that god sirf one is and complete universe among sirf that only pardes (worship) of below is , and universe among each and every thing , sexiest and began , view and invisible his desire of front ndf and quiet is . islam religion of holy book ka name koran is whose hindi among means guttural path is . its followers ko arabic among muslim who go is , whose verbs muslim happens is . muslim yh undead there is that quran reciters (crucifixion among pen) took one of frisbee dhara , muhammad saheb ko nagari century of arab man , nearly 23 year among yad-kanthus tha online made tha . muslim islam ko but this no religion feel . his like god ne muhammad saheb from before also earth on many dut is sent , in which suras , genesis and middle there is . muslims of being genesis and middle of many lectures ko people ne folklore do had . maximum muslims of for muhammad saheb god of shunga dut and the koran manus race of for shunga message is .

Small Bitext Translation (6)

islam religion ( ) الإسلام। after the christian religion on the basis of followers to the world 's second largest religion . islam is the word of the arabic language is the word , which is the word salma of which are two means अमन (1) (2) submission and peace .
ईस्लाम believes in monotheism . its followers believe that god is the only one in the whole universe , and only the glory of the (लायक ,) , and all the things in the creation and बेजान चीज़ , ज़िंदा , god , are at his will and quiet . islam 's holy book is the name of the quran , which means recitation in hindi . its followers are called muslims in arabic , is the plural of muslims . muslims believe that quran in christianity) gabriel (gabriel , by the faristha named to prophet muhammad in the 7th century in arabia , almost 23 years in the verses of the याद-कंठस् was made to islam . muslims do not believe that there is no new religion . according to him , before the prophet muhammad , god has also sent many messengers on earth , in which includes , musa and isa are included . according to the muslims , musa and isa many people pervert . for the majority of muslims mohammed saheb was the last prophet of god and quran for the human race is the last message .

Small Bitext Trans + Dict + Translit + Induced + Mono Scoring (7)

islam , christianity) alevi (religion after the followers on the basis of the world is the second largest religion islam . word of the arabic language from which means is are (1) aman and peace (2) surrender .
islam monotheism . this is to the head of the followers believe that there is only one god , and only in the whole world , he is the lack of) worship (, and in every thing , sexiest creation and imran , god , in front of the wish and are peaceful . islam 's holy book is the name of the quran recitation in hindi , which means its adherents . muslims in arabic , it is said , whose big muslim . muslims believe that quran gabriel (christianity in gabriel) named a frisbee by the , to the prophet muhammad in the 7th century , the arab almost 23 years in the verses of the yad-kanthus . muslims , islam is a new religion . do not believe according to mohammed before god has also sent envoy on earth , in which there are many including , musa bc . muslims are included , and according to the teachings of musa bc , and many people to deform . many muslims for the messenger of god 's prophet muhammad was the last of the quran and the last message for mankind .

Reference Human Translation (8)

Islam is the second largest religion in the world after Christianity, based on the number of followers. Islam is an Arabic word from the root word Sallama, which has two definitions 1) peace and harmony 2) surrender. Islam is a monotheistic religion. The primary belief of its followers is that there is one God and That alone is worthy of worship and all animate and inanimate, visible and invisible objects in nature surrender peacefully to Its will. The holy book of Islam is called the Quran, which in Hindi means vocal chant.A follower is called Muslim in Arabic, the plural of which is Muslims. Muslims believe that Quran was learnt by Mohammad in 7th century Arabia, from Jibrail (Gabriel in Christianity) over approximately 23 years. The Muslims do not believe Islam to be a new religion. According to them, God has sent many messengers on Earth before Mohammad, which includes Abraham, Moses and Jesus. According to Muslims, many sermons of Moses and Jesus have been distorted by people. For most Muslims, Mohammad was the last messenger of God and Quran is the last message for mankind.

Fig. 10: First paragraph of Hindi Wikipedia page on *Islam*, and a progression of translations of it.

References

Alfonseca, Enrique, Massimiliano Ciaramita, and Keith Hall. 2009. Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85, June.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Cherry, Colin 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)*.

Chu, Chenhui, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 8404 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 296–309.

Church, Kenneth W. and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1:163–190, 6.

Church, Kenneth W. and William A. Gale. 1999. Inverse document frequency (IDF): A measure of deviations from Poisson. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*. Springer Netherlands, pages 283–295.

Church, Kenneth W and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Daumé, Hal 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)*.

Dou, Qing and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1676, Seattle, Washington, USA, October. Association for Computational Linguistics.

Dou, Qing, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 557–565, Doha, Qatar, October. Association for Computational Linguistics.

Fung, Pascale. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Workshop on Very Large Corpora*.

Fung, Pascale 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)*.

Hermjakob, Ulf, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of ACL-08: HLT*, pages 389–397, Columbus, Ohio, June. Association for Computational Linguistics.

Irvine, Ann. 2014. *Using Comparable Corpora to Augment Low Resource SMT Models*. Ph.D. thesis, Johns Hopkins University, Department of Computer Science, Baltimore, Maryland.

Irvine, Ann and Chris Callison-Burch. 2013a. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Irvine, Ann and Chris Callison-Burch. 2013b. 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)*.

Irvine, Ann and Chris Callison-Burch. In submission. Discriminative bilingual lexicon induction. *Computational Linguistics*, pages 1–42.

Irvine, Ann, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Klementiev, Alex, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Klementiev, Alexandre 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)*.

Koehn, Philipp, 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)*.

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

Li, Haizhou, A Kumaran, Vladimir Pervouchine, and Min Zhang. 2009. Report of news 2009 machine transliteration shared task. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 1–18, Suntec, Singapore, August. Association for Computational Linguistics.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics (TACL)*, 2(January).

Peirsman, Yves and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California, June. Association for Computational Linguistics.

Pekar, Viktor, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*.

Pierrehumbert, Janet B. 2012. Burstiness of verbs and derived nouns. In Diana Santos, Krister Lindén, and Wanjiku Nganga, editors, *Shall We Play the Festschrift Game?* Springer Berlin Heidelberg, pages 99–115.

Post, Matt, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

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

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

Ravi, Sujith and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37:141–188.

Virga, Paola and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 57–64, Sapporo, Japan, July. Association for Computational Linguistics.

Vulić, Ivan, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479–484, Portland, Oregon, USA, June. Association for Computational Linguistics.

Vulić, Ivan and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA, October. Association for Computational Linguistics.