A Comprehensive Analysis of Bilingual Lexicon Induction

Ann Irvine*
Johns Hopkins University

Chris Callison-Burch**
University of Pennsylvania

Bilingual lexicon induction is the task of inducing word translations from monolingual corpora in two languages. In this paper we present the most comprehensive analysis of bilingual lexicon induction to date. We present experiments on a wide range of languages and data sizes. We examine translation into English from 25 foreign languages: Albanian, Azeri, Bengali, Bosnian, Bulgarian, Cebuano, Gujarati, Hindi, Hungarian, Indonesian, Latvian, Nepali, Romanian, Serbian, Slovak, Somali, Spanish, Swedish, Tamil, Telugu, Turkish, Ukrainian, Uzbek, Vietnamese and Welsh. We analyze the behavior of bilingual lexicon induction on low frequency words, rather than testing solely on high frequency words, as previous research has done. Low frequency words are more relevant to statistical machine translation, where systems typically lack translations of rare words that fall outside of their training data. We systematically explore a wide range of features and phenomena that affect the quality of the translations discovered by bilingual lexicon induction. We give illustrative examples of the highest ranking translations for orthogonal signals of translation equivalence like contextual similarity and temporal similarity. We analyze the effects of frequency and burstiness, and the sizes of the seed bilingual dictionaries and the monolingual training corpora. Additionally, we introduce a novel discriminative approach to bilingual lexicon induction. Our discriminative model is capable of combining a wide variety of features, which individually provide only weak indications of translation equivalence. When feature weights are discriminatively set, these signals produce dramatically higher translation quality than previous approaches that combined signals in an unsupervised fashion (e.g. using minimum reciprocal rank). We also directly compare our model's performance against a sophisticated generative approach, the matching canonical correlation analysis (MCCA) algorithm used by Haghighi et al. (2008). Our algorithm achieves an accuracy of 42% versus MCCA's 15%.

1. Introduction

In natural language processing, translations are typically learned from parallel corpora, which are sentence-aligned bilingual texts (Brown et al. 1990). In contrast, bilingual lexicon induction is the task of inducing word translations from monolingual corpora in two languages. These monolingual corpora can range from being completely unrelated topics to being comparable corpora that contain related information (like Wikipedia articles on the same subject, but written independently in two languages), but they are not translations of each other. Being able to learn translations from monolingual text is potentially very useful for machine translation (MT). For many language pairs, we often

* Center for Language and Speech Processing, 3400 N Charles Street Baltimore, MD 21218. E-mail: annirvine@gmail.com
** Computer and Information Science Department, 3330 Walnut Street, Philadelphia, PA 19104. E-mail: ccb@upenn.edu

only have access to small bilingual resources. When a machine translation system has access to limited parallel corpora and to incomplete bilingual dictionaries, therefore, there are likely to be many unknown (out-of-vocabulary, or OOV) words in the texts that we would like it to translate. Being able to mine translations for these OOV words from monolingual corpora means that we could potentially produce some translation for every word in our text, achieving perfect model coverage (but not perfect accuracy).

Bilingual lexicon induction uses monolingual or comparable corpora to identify pairs of translated words. Additionally, a small seed dictionary is also typically assumed. The quality of induced word translations could be evaluated by using the induction algorithm to expand the coverage of translation models extracted from parallel corpora, by translating OOV words, and then checking whether the induced translations improved the MT system. However, most prior work in bilingual lexicon induction has treated it as a standalone task, without actually integrating induced translations into end-to-end machine translation. Instead, it has been evaluated by holding out a portion of the bilingual dictionary and evaluating how well the algorithm learns the translations of the held out words.

To discover translated words across languages, past work has proposed a variety of monolingual distributional similarity metrics as signals of translation equivalence. These signals include contextual similarity, temporal similarity, and orthographic similarity. Most prior work has used unsupervised methods (like rank combination) to aggregate these types of orthogonal signals (Schafer and Yarowsky 2002; Klementiev and Roth 2006). Surprisingly, no past research has employed *supervised* approaches to combine diverse monolingually-derived signals for bilingual lexicon induction. The field of machine learning has shown repeatedly that supervised models dramatically outperform unsupervised models, including for closely related problems like statistical machine translation (Och and Ney 2002). For the bilingual lexicon induction task, a supervised approach is natural, particularly because computing contextual similarity typically requires a seed bilingual dictionary (Rapp 1995), and that same dictionary may be used for estimating the parameters of a model to combine monolingual signals. In this setting, bilingual lexicon induction is critical for translating source words which do not appear in the parallel data or dictionary.

We make several contributions with this article.[1] First, we present a discriminative model of bilingual lexicon induction that significantly outperforms previous models. Our discriminative model is capable of combining a wide variety of features, which individually provide only weak indications of translation equivalence. When feature weights are discriminatively set, these signals produce dramatically higher translation quality than previous approaches that combined signals in an unsupervised fashion (e.g. using minimum reciprocal rank). We present experiments results showing consistent improvements in translation accuracy for 25 languages. The absolute accuracy increases over the MRR baseline ranges from 5% to 31%, which correspond to 36% to 216% relative improvements. Moreover, we directly compare our model's performance against a sophisticated generative approach, the matching canonical correlation analysis (MCCA) algorithm used by Haghighi et al. (2008). Our algorithm achieves an accuracy of 42% versus MCCA's 15%, again showing the advantages of our discriminative approach.

Second, our experimental settings represent more realistic and more useful settings than those used by previous work. Previous work in bilingual lexicon induction only

1 This article expands research previously published in Irvine and Callison-Burch (2013) and Irvine (2014).

reports results on inducing translations for the most frequent source language words, completely avoiding any scalability or data sparsity issues. Because those word counts are not sparse, that task is much easier than inducing translations for a randomly drawn set of words. We analyze the accuracy of our algorithm in terms of the frequency of words, in order to understand the effects of data sparseness. Previous work frequently simulates low-resource languages, often focusing on Spanish-English or German-English translation and limiting the large resources available for those languages. We present experimental results on a wide variety of languages, for which a wide variety of monolingual corpora and seed bilingual dictionaries are available. Many of our languages are genuinely low-resource.

Third, we systematically explore a wide range of features and phenomena that affect the quality of the translations discovered by bilingual lexicon induction. We give illustrative examples of the highest ranking translations for orthogonal signals of translation equivalence, including contextual similarity, temporal similarity, orthographic similarity, and topical similarity. We analyze the effects of frequency and burstiness, and the sizes of the seed bilingual dictionaries and the monolingual training corpora. We calculate the correlation between our different signals of translation equivalence, in order to quantify how orthogonal they are. We present an analysis of how accurate each signal is based on the part of speech of the words being translated.

This article represents the most comprehensive investigation into bilingual lexicon induction to date.

2. Monolingual Signals of Translation Equivalence

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. For a given source language word, we score all target language candidates separately and then rank them. We use a variety of signals derived from source and target monolingual corpora as features and use supervision to estimate the strength of each. A diverse range of signals have been used for bilingual lexicon induction in past work, notably by Rapp (1995), Fung (1995), Schafer and Yarowsky (2002), Klementiev and Roth (2006), Klementiev et al. (2012), and others. In this section, we detail the signals of translation equivalence that we use as components in our discriminative model.

2.1 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 Pantel 2010), context vector representations can also be used to compare the similarity of words across two languages. The earliest work in bilingual lexicon induction by Rapp (1995) and Fung (1995) used the surrounding context of a given word as a clue to its translation.

The key to using contextual similarity as a signal of translation equivalence is to find a mapping between the vector space of one language and the vector space of another language. To accomplish this, Rapp (1995) originally proposed creating two co-occurrence matrices for the source and target languages, where the co-occurrence between a pair of words is defined as follows:

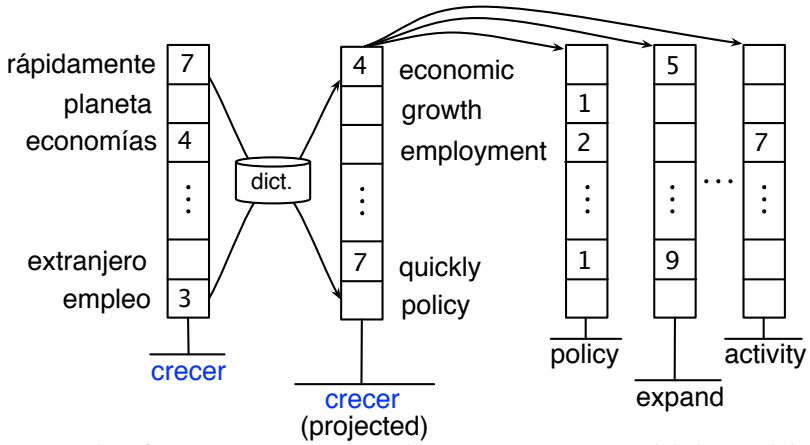


Figure 1: Example of projecting contextual vectors over a seed bilingual lexicon. In monolingual text, Spanish *crecer* appears in the context of the words *empleo, extranjero*, etc. A context vector is built and projected across a seed dictionary. Context vectors for English words (*policy, expand*, etc.) are collected and then compared against the projected context vector for Spanish *crecer* (which can be glossed as *grow*). Words with similar context vectors are likely to be translations of one another.

$$A_{i,j} = \frac{(f(i,j))^2}{f(i) \cdot f(j)}$$

Where $f(i,j)$ is defined as the number of times words i and j, in the same language, occur in the same context in a large monolingual corpus (Rapp (1995) uses a context window of 11 words), and $f(i)$ is the total number of times word i appears in the same corpus. In this original formulation, no bilingual information was employed to find the mappings between the vector spaces of the two languages. Instead, after computing the two co-occurrence matrices for the two languages, Rapp (1995) iteratively randomly permutes the word order of the matrix for one of the languages and calculates the similarity between the two matrices. The permutation is optimal when the similarity between the matrices is maximal, which is when the ordered words in the two matrices are most likely to be translations of one another. Results are given for a set of 100 English and German word translation pairs.

Later formulations of the problem, including Fung and Yee (1998) and Rapp (1999), 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. That is, each position in contextual vector v corresponds to a word in the source vocabulary[2], and vectors v are computed for each source word in the test set. Fung and Yee (1998) calculates the *i*th position of word w's context vector, v_{w_i}, as:

$$v_{w_i} = TF_{i,w} \cdot IDF_i$$

2 In fact, they need only correspond to those source words which have translations in the seed bilingual dictionary.

alcanzaron	sanitario	desarrollos	volcánica	montana
reached	exil	advances	**volcanic**	arendt
enjoyed	rhombohedral	**developments**	eruptive	**montana**
contained	apt	changes	coney	glasse
contains	immune	placing	rhonde	teter
saw	circulatory	innovations	bleaker	waddingham
includes	nervous	use	staten	daryl
included	endocrine	changes	robben	callowhill
hit	coordinate	making	ostrov	richings
achieved	ucsd	addition	ellesmere	beswick
estates	windowing	allowing	gilligan	holgersson

Table 1: Examples of translation candidates ranked using contextual similarity. The correct English translations, when found, are bolded. English words are ordered by their contextual similarity scores with the given Spanish word. Here are glosses of the Spanish words: *alcanzaron: reach*, *sanitario: sanitary*, *desarrollos: development/growth*, *volcánica: volcanic*, and *montana: montana*.

Where $TF_{i,w}$ is the number of times i and w co-occur (in this case, defined as appearing in the same sentence), and:

$$IDF_i = log\frac{maxn}{f_i} + 1$$

Where $maxn$ is the maximum frequency of any of the words in the corpus, and f_i is the frequency of word i. Rapp (1999) uses log-likelihood ratios instead of $TF \cdot IDF$. Once source and target language contextual vectors are built, each position in the source language vectors is projected onto the target side using a seed bilingual dictionary.[3] Finally, *contextual similarities* are calculated. That is, each projected vector is compared, using any vector comparison method, with the context vector of each target word. Word pairs with high contextual similarity are likely to be translations. This method of projecting contextual vectors is illustrated in Figure 1. Rapp (1999) uses the same projection method as Fung and Yee (1998) but uses log-likelihood ratios instead of $TF \cdot IDF$.

We use the vector space approach of Rapp (1999) 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 1). 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 f is projected onto the English vector space using translations in a given bilingual dictionary to map the component values into their appropriate English

3 This is the only time that the bilingual dictionary was used, except for evaluation. In our approach, we also use the seed bilingual dictionary as supervision for a discriminative model.

alcanzaron	sanitario	desarrollos	volcánica	montana
travel	snowpocalypse	occupied	wawel	dzv
road	airport	aer	**volcanic**	spatz
news	dioxide	madoff	ash	centimes
services	steinmeier	declaration	spewed	kleve
arts	gobbling	ponzi	eyjafjallajokull	reallocate
word	investigating	affects	otunbajewa	frostrup
special	convicted	suspected	eruption	roze
chief	spy	fed	cloud	minc
top	offices	combat	rubell	bicyclists
inspired	bond	arrested	dormancy	lgbt

Table 2: Examples of translation candidates ranked using temporal similarity. The correct English translations, when found, are bolded. English words are ordered by their temporal similarity scores with the given Spanish word.

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 $c(f, e)$ 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 Yee (1998), Rapp (1999)). Following Fung and Yee (1998), we compute the value of the k-th component of f's contextual vector, f_k, as follows:

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

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, respectively, 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:

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

Table 1 shows example ranked lists using contextual similarity to rank English words for several Spanish words. For example, contextual similarity ranks the English words *enjoyed*, and *contained* highly as candidate translations of Spanish *alcanzaron*. These incorrect English words tend to appear in similar contexts as the correct English translation, *reached*.

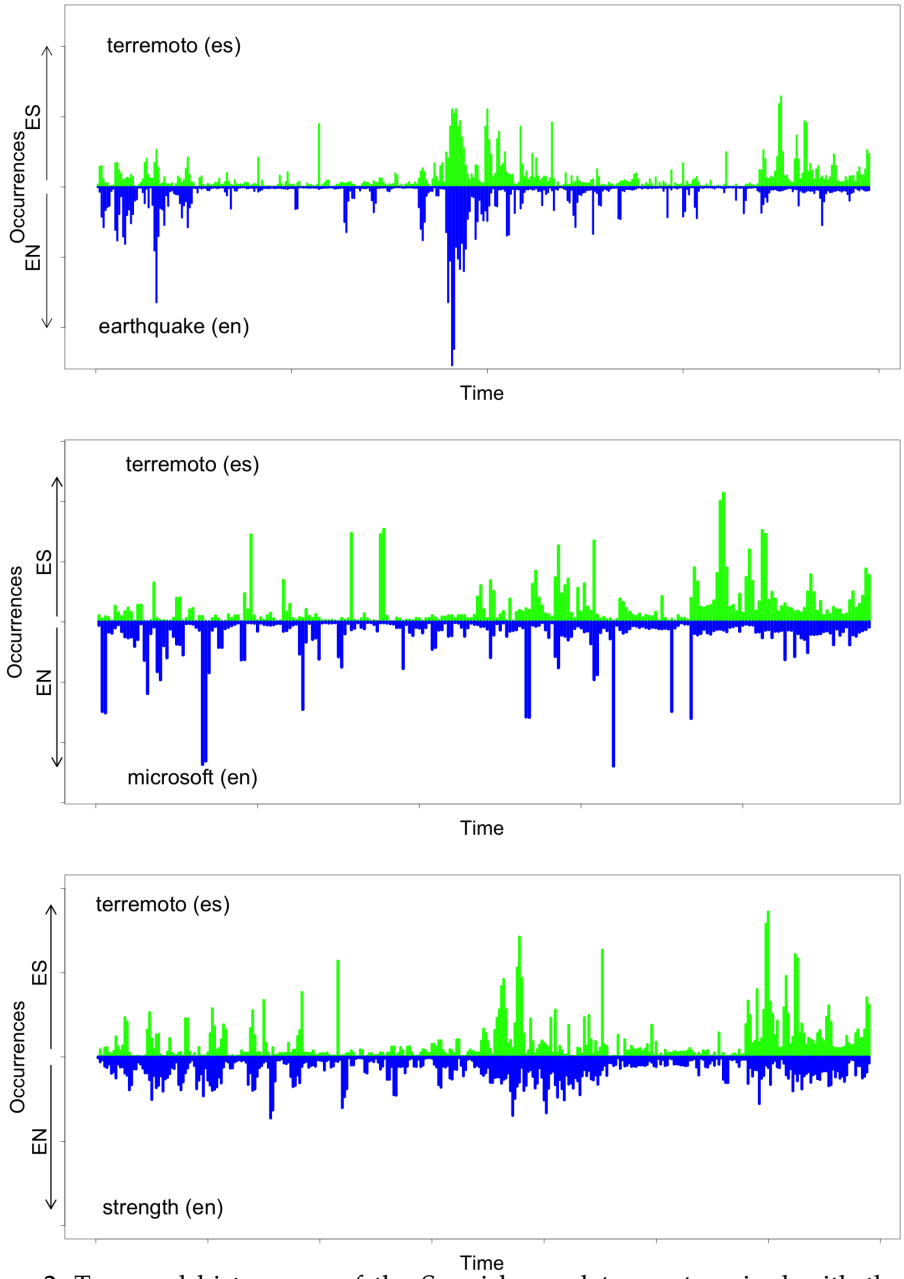


Figure 2: Temporal histograms of the Spanish word *terremoto* paired with three English candidate translations: the correct translation *earthquake* and the incorrect candidates *microsoft* and *strength*. The temporal histograms are collected from monolingual texts spanning several years and show the number of occurrences of each word (on the y-axes) across time. While the correct translation has a good temporal match (sim_{temp}(terremoto, earthquake) $= 2 \cdot 10^{-4}$), the non-translations are less temporally similar (sim_{temp}(terremoto, microsoft) $= 2 \cdot 10^{-5}$, sim_{temp}(terremoto, strength) $= 3 \cdot 10^{-5}$). *In all examples, only dimensions (dates) which are non-zero valued for both signatures are shown, which results in the signature for terremoto appearing somewhat different across the three comparisons.*

2.2 Temporal Similarity

Usage of words over time may be another signal of translation equivalence. The intuition 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. Figure 2 illustrates how the temporal distribution of Spanish *terremoto* is more similar to its English translation *earthquake* than to other English words. *Microsoft*, one of the non-translations, like *earthquake*, is very bursty (formal definition given in Section 2.6). *Strength*, another non-translation, in contrast, appears with fairly consistent frequency over time. The temporal histograms for *terremoto* and *earthquake* both show significant peaks in the middle of the series, which correspond to the major earthquake that occurred in Haiti in January of 2010. Although the two words have reasonably well matched temporal signature, there are some differences. For example, a small earthquake in South America might be covered in Spanish news but not in English news. Other things have periodic temporal signatures, like words associated with the Olympics, the World Cup or the US presidential election.

To calculate temporal similarity, we collected online monolingual newswire over a multi-year period and associate each article with a time stamp. Each document in our web crawls of online news websites has an associated publication date (see Section 3.3). 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 Yarowsky (2002), Klementiev and Roth (2006), Alfonseca, Ciaramita, and Hall (2009).

We calculate $sim_{temp}(F_{temp}, E_{temp})$, the temporal similarity between a pair of words, using the method defined by Klementiev and Roth (2006). We generate a temporal signature for each word by sorting the set of (time-stamped) documents in the monolingual corpus into a sequence of equally sized temporal bins and then counting the number of word occurrences in each bin. In our experiments, our English web crawl data is vastly outstrips the other languages, so we restrict the English data that we use in a particular foreign language experiment to be no more than three times the size of our source language web crawled data, and only include news articles from those dates for which we also have source language articles. We again use cosine similarity to compare the normalized temporal signatures for a pair of words:

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

where F_{temp} and E_{temp} are source and target language word temporal signatures, respectively. The k-th component of a word f's temporal vector, f_k, represents the frequency of the word f during the k-th date range in the temporal bins created for the time-stamped monolingual corpora. The size of the two vectors used for temporal similarity calculation is a function of the number of temporal bins. In our experiments, we set the temporal bin size to 3 days, so the size of temporal signatures is equal to the number of days spanned by a monolingual corpus divided by three. We normalize the temporal signature of each word by dividing all of f_k components by the total count of the word f. In Irvine (2014), we compared the performance of using raw temporal

alcanzaron	sanitario	desarrollos	volcánica	montana
alcantara	**sanitary**	ferroalloy	**volcanic**	**montana**
albanian	sanitation	barrosos	volcanism	fontana
lazzaroni	unitario	destroyers	voltaic	montane
lanaro	sanitarium	mccarroll	vacancy	mentana
aleandro	sanitation	disallows	konica	montagna
lazaros	sagittario	disallow	dominica	montanha
canaro	sanitarias	scrolls	veronica	montan
alianza	kantaro	payrolls	monica	montano
lazaro	sanitorium	carroll	volcano	montani
catanzaro	santoro	steamrolls	vratnica	montand

Table 3: Examples of translation candidates ranked using orthographic similarity. The correct English translations, when found, are bolded. English words are ordered by their orthographic similarity scores with the given Spanish word.

signatures and using the Discrete Fourier Transform of those signatures, and found that raw temporal signatures performed just as well as DFT signatures.

Table 2 shows example ranked lists using temporal similarity to rank English words for several Spanish words. For example, *ash* and *spewed*, as well as the Icelandic volcano *eyjafjallajokull*, are all temporally similar to the Spanish word *volcánico*. Since volcanic eruptions are dramatic events that are usually written about in newspapers all around the world when they occur, it is not surprising that this signal is able to produce a correct translation for *volcánico*, alongside highly ranking several related words.

2.3 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 similarity between a pair of words. We use the edit distance between the two words, normalized by the average of the lengths of the two words:

$$sim_{orth}(f, e) = \frac{ed(f, e)}{\frac{|e|+|f|}{2}}$$

where *ed* is the standard Levenshtein edit distance between the two strings. This is straightforward for languages which use the same character set, but it is more complicated for languages that are written using different scripts. A variety of prior work has focused on the problem of learning mappings between character sets (e.g. Yamada and Knight (1999), Tao et al. (2006), Yoon, Kim, and Sproat (2007), Bergsma and Kondrak (2007), Li et al. (2009), Snyder, Barzilay, and Knight (2010), Berg-Kirkpatrick and Klein (2011)).

For non-Roman script languages, we transliterate words into the Roman script before measuring orthographic similarity with their candidate English translations. Following prior work (Virga and Khudanpur 2003; Irvine, Callison-Burch, and Klementiev 2010), we treat transliteration as a monotone character translation task and train models on the mined pairs of person names in foreign, non-Roman script languages and English. Our MT-based transliteration system can translate a single character as many

alcanzaron	sanitario	desarrollos	volcánica	montana
reached	health	**developments**	**volcanic**	**montana**
began	transcultural	developed	eruptions	miley
led	medical	development	volcanism	hannah
however	sanitation	used	lava	beartooth
early	patient	using	plumes	cyrus
including	deliverables	modern	eruption	crazier
took	pharmaceutical	based	volcano	bozeman
remained	sewerage	important	volcanoes	chelsom
several	healthcare	history	breakouts	absaroka
continued	care	different	volcanically	baucus

Table 4: Examples of translation candidates ranked using topic similarity. The correct English translations, when found, are bolded. English words are ordered by their topic similarity scores with the given Spanish word.

characters, and it can translate multiple input characters into a single output character. Because transliteration is strictly a monotone operation, we do not allow reordering in our models. Additionally, unlike in machine translation, our translation and language models can support very large n-gram sizes because the number of characters in a given script is small compared to word vocabularies; we use phrase length limits of 10 when extracting translation grammars and in estimating language models. We use a character-based language model trained on a list of English names.

In Irvine, Callison-Burch, and Klementiev (2010), we provide a detailed evaluation of our transliteration technique, and found it to be competitive with the best performing system in a transliteration shared task (Li et al. 2009). For purposes of bilingual lexicon induction, we use the top-1 transliteration to compute edit distance.

Table 3 shows example ranked lists using orthographic similarity to rank English words for several Spanish words. For those Spanish words that have English cognates, such as *sanitario* and *volcánica*, the orthographic signal ranks correct translations highly. For Spanish words without English cognates, like *desarrollos* or *alcanzaron*, the English words with the highest orthographic similarity are unrelated to the Spanish words.

2.4 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. If we were able to associate articles about the same topic across two languages, then we ought to be able to use that to compute a topic similarity score to help rank potential translations. We Wikipedia articles create topic signatures for words. Figure 3 illustrates this idea. The figure shows a topic vector for the English word *troops* and 3 Russian words. The counts in the vector for *troops* are the number of time that it occurred in the Wikipedia article corresponding to that position in the vector. For instance, the word *troops* occurred 15 times on the Wikipedia article about *Barack Obama*. How can we associate topics across languages? In order to find a mapping of topics across languages, we use Wikipedia's interlingual links, in a similar fashion that we used the small seed bilingual dictionaries to project across the vector spaces for two languages when computing contextual similarity.

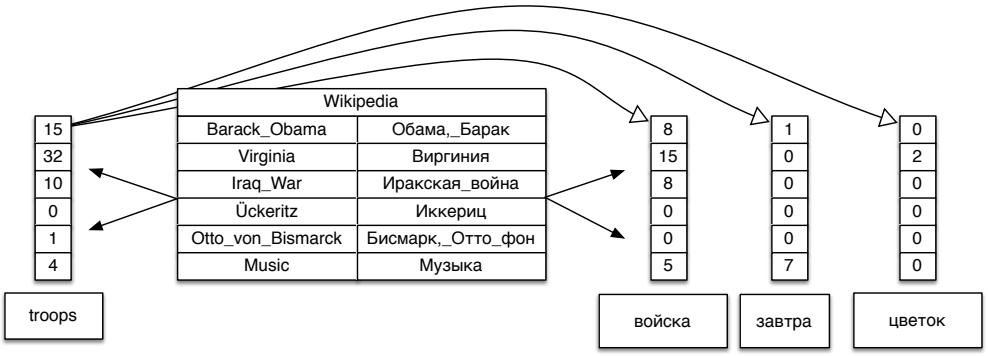


Figure 3: Illustration of how we compute the topical similarity between *troops* and three Russian candidate translations. We first collect the topical signatures for each word (e.g. *troops* appears in the page about *Barack Obama* 15 times and in the page about *Virginia* 32 times.) based on the interlingually linked pages. We can then directly compare each pair of topical signatures. English glosses for the three Russian words are (from left to right): *troops, tomorrow* and *flower*.

In order to score how likely a pair of words f and e are to be translations, we compare their topic signatures F and E, by counting the the words' occurrences in each topic, normalize the signatures, and then comparing the resulting vectors. We simply compute cosine distance between topic signatures.

$$sim_{topic}(F_{topic}, E_{topic}) = \frac{F_{topic} \cdot E_{topic}}{||F_{topic}||||E_{topic}||}, \tag{4}$$

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 For each foreign language, the number of Wikipedia articles linked to English pages is given in Table 6. 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).

Table 4 shows examples of English words ranked using topic similarity for several Spanish words. Using topic similarity, *montana*, *miley*, *cyrus* 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. Likewise, Bozeman is a large city in Montana, and Max Baucus represented the state in the US Senate for over 35 years.

2.5 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,

Frequency, f, and number of words, n	IDF		Burstiness	
	Top-5	Bottom-5	Top-5	Bottom-5
$f = 50, n = 802$	kratsa tebet kagome khaldūn psittacosaurus	contemporaneously unrecognizable categorizing modern-style crazed	straubing-bogen tebet cloppenburg autosan gøta	wavering busing unconvinced redesigning oftentimes
$f = 100, n = 303$	subarticle trackmania lyrebird gârbea biecz	call-ups workable purports outnumber unmatched	penedès lyrebird azarbaijan padstow trackmania	demoralized misgivings precluded workable forestall

Table 5: Examples of highest and lowest ranked English words according to two measures of burstiness. Empirical estimates were taken from a subset of English Wikipedia data.

or:

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

This helps prevent high frequency closed class words from being considered viable translations of less frequent open class words.

2.6 Burstiness Similarity

Burstiness is a measure of how peaked a word's usage is over a particular corpus of documents (Pierrehumbert 2012). Bursty words are topical words that tend to appear frequently in a document when some topic is discussed, but do not not frequently across all documents in a collection. 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 Gale (1995, 1999) provide an overview of several ways to measure burstiness empirically. Following Schafer and Yarowsky (2002), we measure the burstiness of a given word in two ways. The first is based on Inverse Document Frequency (IDF):

$$IDF_w = -log\frac{df_w}{|D|}, \tag{5}$$

where df_w is the number of documents that w appears in, and $|D|$ is the total number of documents in the collection. The second burstiness measure, similar to that defined by Church and Gale (1995), is 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 Gale (1995) and use relative frequencies rather than absolute frequencies to account for varying document lengths.

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

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. Table 5 shows examples of high and low ranked bursty words under each measure for two different constant word frequencies. The examples show that both measures of burstiness yield rankings that are consistent with our intuitions, yet they provide different results.

We compare both the IDF and the B scores for pairs of words using ratios:

$$sim_{IDF}(e, f) = min[\frac{IDF_e}{IDF_f}, \frac{IDF_f}{IDF_e}]$$

$$sim_{burst}(e, f) = min[\frac{B_e}{B_f}, \frac{B_f}{B_e}]$$

2.7 Variations and Additional Signals

We perform experiments using variations on the signals listed above. Two variations are word prefix contextual similarity and word suffix contextual similarity. Prefix contextual similarity is calculated in the same way as the contextual similarity score, but we use source and target word stems, or word prefixes up to five characters long, instead of full words. That is, the word prefix contextual similarity score for the word pair (*blanc*o, *white*) is the same as that of (*blanc*a, *white*). In this particular example, we collect only a single contextual vector for *blanc{o,a}*. In Spanish, this translation of the English word *white* appears with either a masculine or feminine ending, depending on what it modifies. By summing the distributional counts of *blanco* and *blanca*, we expect a contextual vector that is more similar to English *white* than either alone. We measure the similarity of a pair of prefixal contextual vectors using cosine similarity, as before.

Suffix contextual similarity measure is similar to the word stem measure, except instead of using word prefixes, it uses word *suffixes* of up to five characters long. For example, the word stem contextual similarity score of the word pair (impo*sible*, pos*sible*) is the same as that of (po*sible*, impose*sible*). With this signal, we expect to sum over alternate word prefixes in the same way that the word stem signal sums over alternate word suffixes. The intuition is that suffix similarity may help to group words with the same syntactic classes. Again, the similarity between a pair of suffixal contextual vectors is measured using cosine similarity. In addition to prefix and suffix contextual similarity, we also estimate prefix and suffix topic and temporal similarity.

We also use an indicator feature which is positive if the source and target words are the same string. Of course, this indicator is most useful for languages written in the same script.

Finally, we add a final feature indicating the target translation's monolingual frequency, which serves as a sort of prior probability that the target word is of interest at all. Specifically, we define this feature as the inverse of the log of the target word's frequency.

Although we have limited our experiments to this set of varied signals of translation equivalence, our basic framework is easily extendible.

3. Experimental Setup

We designed a set of experiments to systematically explore the following research questions: To what extent are the different signals of translation equivalence orthogonal

to each other? Are certain signals better than others at ranking translations? Does this vary based on language or part of speech? How accurately do they individually rank translation candidates for a variety of languages? How can we effectively combine them in order to rank translation candidates? How much does the performance vary per language? To what extent does performance depend on the size of the seed bilingual dictionary, and on the size of the monolingual corpora? Does bilingual lexicon induction make more accurate predictions for words with certain properties like being highly bursty? How well does our discriminative model compare to the sophisticated generative model MCCA?

First, we describe our evaluation metric, data, and experimental setup. Then we present our findings.

3.1 Evaluation metric

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:

$$acc_k = \frac{\sum_{l \in L} I_{lk}}{|L|} \tag{7}$$

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.

3.2 Bilingual dictionaries

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 HITs asking workers to translate the most frequent $10,000$ words in the most viewed $1,000$ pages for each source language. All of the source words in the Wikipedia dictionaries are unigrams, we allowed workers to translate them into multi-word English phrases, but we only used entries that were translated as single words for the experiments described in this article. Workers were shown words in the context of three Wikipedia sentences. Additional details on experimental design and quality control mechanisms are given in Pavlick et al. (2014). As a result of that project, we collected bilingual dictionaries of about $10,000$ words translated into English. For the experiments in this article, we filter the dictionaries to include only high quality translations. Specifically, we only use translations that have a quality score of at least 0.6 under the worker quality metric given by Pavlick et al. (2014).

Language	Dict entries (freq >= 10)	Wikipedia words	interlanguage links	Web crawl words	Web crawl dates
Albanian	7,314	6,388,669	19,860	9,127,415	598
Azeri	5,668	6,747,026	26,896	3,842,179	176
Bengali	5,368	4,998,454	18,603	8,295,164	467
Bosnian	7,139	7,515,961	19,981	8,647,129	794
Bulgarian	8,587	33,926,577	88,436	34,042,882	1208
Cebuano	899	2,755,209	52,026	1,886,463	121
Gujarati	4,442	3,958,031	3,909	1,084,719	122
Hindi	6,585	16,198,183	25,078	31,123,091	823
Hungarian	2,268	69,695,400	127,406	542,736	119
Indonesian	4,805	26,769,690	83,274	5,067,534	623
Latvian	7,311	9,432,914	33,024	36,156,391	747
Nepali	3,535	1,878,168	5,854	3,489,101	179
Romanian	6,600	34,672,327	135,874	17,608,197	374
Serbian	7,403	37,575,834	131,854	15,194,828	550
Slovak	7,346	23,477,764	107,958	113,163,058	1043
Somali	1,125	267,383	1,470	3,250,014	322
Spanish	7,780	232,437,776	374,651	913,465,084	3718
Swedish	5,534	70,923,386	274,152	11,307,825	122
Tamil	4,735	9,154,660	23,468	3,928,554	157
Telugu	5,136	8,769,259	8,841	3,254,373	120
Turkish	6,139	30,385,844	89,577	14,409,942	1165
Ukrainian	8,469	72,135,536	208,915	21,836,916	1350
Uzbek	969	5,368,879	71,081	8,304,074	333
Vietnamese	1,823	53,471,136	194,374	2,468,179	121
Welsh	4,207	4,414,153	28,066	6,573,628	704
Average	5,247	30,932,729	86,185	51,122,779	635
Median	5,534	9,432,914	52,026	8,304,074	467

Table 6: Statistics about the data used in our experiments.

3.3 Monolingual Data

We draw monolingual data from two sources: (1) web crawls of online newspapers, and (2) Wikipedia. Table 6 gives stats about the amount of data that we gathered for each language.

3.3.1 Web crawls. Online newspapers are good sources of text for many languages. We began harvesting such data by crawling several well-known news sources that publish stories in two or more languages, including Deutsch Welle and Voice of America. In order to gather more data, particularly for less commonly used languages, we scraped a list of $44,892$ newspapers and their locations, URLs, and languages from the ABYZ News Links website.[4] The resulting database of newspapers contains links to online newspapers published in 128 languages, and we set up web crawls to download the content from each daily.

Because our data is comprised of news stories, each document also has an associated time stamp, which we use to define a rough document alignment with English news articles. That is, we treat the set of all foreign language news stories published on a particular day as roughly comparable to those written in English on the same day. The degree of comparability between such sets of documents varies greatly.

4 www.abyznewslinks.com/

3.3.2 Wikipedia. We also use Wikipedia as a source of monolingual data. For all languages, we use Wikipedia's January 2014 data snapshots. To maximize the degree of comparability between our source language Wikipedia pages and English Wikipedia, we only use those pages which have interlingual links with English pages. Unlike our newspaper web crawls, Wikipedia content has fairly reliable language labels. However, for some languages, English content is copied from the English Wikipedia without translation. We use the CLD2 language ID system to identify and remove English content from other languages' Wikipedias.

We also use Wikipedia as a source for example *transliterations* in non-roman script languages paired with English. In (Irvine, Callison-Burch, and Klementiev 2010), we detailed how we mined transliteration training data from Wikipedia page titles for 150 languages. 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. In Section 2.3, we use this data to train transliterators and transliterate source language words before comparing their orthographies with English words.

3.4 Languages

We report performance results for bilingual lexicon induction English from 24 foreign languages into English. The languages in our study are Albanian, Azeri, Bengali, Bosnian, Bulgarian, Cebuano, Gujarati, Hindi, Hungarian, Indonesian, Latvian, Nepali, Romanian, Serbian, Slovak, Somali, Swedish, Tamil, Telugu, Turkish, Ukrainian, Uzbek, Vietnamese and Welsh. Statistics about the data for each of the languages is given in Table 6.

3.5 Monolingual Signals

In our experiments, we use a total of 18 features to rank English words as potential translations of the input foreign word. These are estimated from our two sources of comparable monolingual data, web crawls and Wikipedia: (1) Web Crawls Contextual Similarity, (2) Web Crawls Temporal Similarity, (3) Orthographic Similarity, (4) Wikipedia Contextual Similarity, (5) Wikipedia Topic Similarity, (6) Wikipedia Frequency Similarity, (7) Wikipedia IDF Similarity, (8) Wikipedia Burstiness Similarity, (9) Web Crawls Prefix Contextual Similarity, (10) Web Crawls Prefix Temporal Similarity, (11) Web Crawls Suffix Contextual Similarity, (12) Web Crawls Suffix Temporal Similarity, (13) Wikipedia Prefix Contextual Similarity, (14) Wikipedia Prefix Topical Similarity, (15) Wikipedia Suffix Contextual Similarity, (16) Wikipedia Suffix Topical Similarity, (17) String Identity, and (18) Inverse Log of Target Wikipedia Frequency.

Table 8 shows examples of the values assigned to several English candidate translations of Romanian words for each of the 18 features.

Language	Candidates	Language	Candidates	Language	Candidates
Albanian	102,998	Hungarian	199,293	Swedish	286,774
Azeri	113,751	Indonesian	157,209	Tamil	89,316
Bengali	76,014	Latvian	115,933	Telugu	54,415
Bosnian	89,871	Nepali	38,895	Turkish	185,906
Bulgarian	181,510	Romanian	203,665	Ukrainian	232,221
Cebuano	59,546	Serbian	188,282	Uzbek	98,191
Gujarati	34,289	Slovak	171,250	Vietnamese	159,240
Hindi	101,777	Somali	43,826	Welsh	97,317

Table 7: Number of candidate English words, by source language. English candidates appear at least ten times in the monolingual corpora.

src	trg	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
politic	**political**	.127	0	0.25	.165	.139	.722	.644	.134	.359	.891	0	0	.465	.179	0	0	0	.095
	offing	0	.879	0.92	0	0	7.414	.391	.027	0	0	0	0	0	0	0	0	0	.402
	first	0	0	1.0	0	.130	2.490	.274	.239	0	0	0	0	0	0	0	.133	0	.081
	shipbuilding	.161	0	0.95	0	0	3.358	.638	.072	0	0	0	0	0	0	0	0	0	.155
curs	**course**	0	0	0.4	0	.055	.437	.820	.036	0	0	0	0	0	.052	0	0	0	.107
	refresher	.092	0	1.08	0	0	7.132	.380	.031	0	0	0	0	0	0	0	0	0	.369
	meeting	.089	0	1.27	0	0	.702	.933	.033	.175	0	0	0	0	0	0	0	0	.110
	pirc	0	0	0.75	0	0	7.374	.358	.038	0	0	0	0	0	0	0	0	0	.402
valea	**valley**	0	.925	0.36	0	0	.036	.693	.184	0	0	0	.919	0	0	0	0	0	.103
	geography	0	0	1.14	0	.012	.074	.509	.377	0	0	0	0	0	0	0	0	0	.102
	either	0	0	0.91	0	.013	.250	.566	.056	0	0	0	0	0	0	0	0	0	.100
	birthday	0	.908	1.08	0	0	1.785	.994	.049	0	0	0	0	0	0	0	0	0	.126
olanda	**netherlands**	.194	0	0.82	.293	0	.218	.805	.247	.349	0	0	0	.315	0	0	0	0	.107
	vows	.121	0	1.2	0	0	3.396	.691	.065	0	0	0	0	0	0	0	0	0	.163
	orava	0	0	0.55	0	0	5.337	.499	.759	0	0	0	0	0	0	0	0	0	.237
	kunduz	0	0	0.83	.235	0	5.415	.471	.688	0	0	0	0	.255	0	0	0	0	.241
revista	**magazine**	0	0	1.07	.208	0	.028	.726	.405	0	0	0	0	.338	.050	.178	.040	0	.105
	takwin	.603	0	1.08	0	0	8.167	0	0	0	0	.061	0	0	0	0	0	0	10
	archeological	.065	0	1.0	0	0	2.832	.771	.373	0	0	0	0	0	0	0	0	0	.149
	hollie	0	0	1.08	0	.047	7.231	.432	.109	0	0	0	0	0	0	0	0	0	.417
adus	**brought**	.398	0	1.09	.260	.091	.311	.630	.428	.329	0	.378	0	0	.091	0	0	0	.104
	centuryfrom	.344	0	1.33	0	0	7.982	0	0	.246	.960	0	0	0	0	0	0	0	10
	associated	0	0	1.29	0	.059	.170	.681	.536	0	.959	0	0	0	.074	0	.062	0	.105
	abuse	0	0	0.44	0	0	1.591	.875	.407	0	0	0	0	0	0	0	0	0	.129

Table 8: Example feature values for Romanian-English word pairs for all 18 features used in our experiments. The feature numbers correspond to those enumerated in Section 3.5. To train our discriminative classifier, we used 1 positive training example and 3 negative training examples. The positive training examples are indicated by English words in bold (dictionary translations). Non-bolded English words are negative training examples (randomly selected word). The values for feature 17 are all 0 since none of the candidate translations are string identical to the input. The values for many other features often round to 0, because they are too low to be shown with 3 significant digits.

3.6 Candidate English Translations

Table 7 shows the number of English words that we consider as candidate translations of the foreign source words for each foreign language. All of these English words are ranked by the 18 monolingual signals for each of the 24 languages.

	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.
wiki-cont	-0.15						
temporal	-0.14	-0.19					
orthography	-0.28	-0.31	-0.28				
topic	-0.15	-0.14	-0.13	-0.30			
frequency	0.01	0.13	0.02	-0.18	0.13		
burstiness	-0.10	0.06	-0.07	0.06	0.11	0.28	
idf	0.06	0.10	-0.12	-0.01	0.00	0.49	0.14

Table 9: Measure of the correlation (orthogonality) between signals. For each of 24 languages, we randomly select $1,000$ source language words and compute the Spearman rank correlation coefficient across pairwise ranked lists of translation candidates generated by each of eight signals of translation equivalence. We average coefficients within each language. The results here show the mean of the correlation coefficient between all pairs of signals across the 24 languages.

4. Analyzing and Combining Signals of Translation Equivalence

In Sections 4.1–4.3 we analyze the strength of our different signals of translation equivalence, and how best to combine them.

4.1 Orthogonality of Signals

The primary goal of this article is to show how a diverse set of weak signals of translation equivalence can be combined to learn the translations of words from monolingual texts. The different signals need to be orthogonal in order for a combination to improve their individual accuracy. Intuitively, the signals that we defined in Section 2 seem to be orthogonal. That is, they provide very different types of information about how words are used in language, and we hypothesize that the lists of ranked candidate translations under each signal are uncorrelated with the exception (and hope!) that correct translation pairs rank relatively high according to all or most of the signals. In our first set of experiments, we measure their orthogonality empirically.

In order to empirically measure orthogonality of our signals, we measure pairwise Spearman rank-order correlation coefficients. Specifically, we first use each signal separately to rank all translation candidates. Then, we measure the correlation between all pairs of ranked lists using the Spearman coefficient. A correlation coefficient of 1.0 indicates perfect positive correlation, -1.0 indicates perfect negative correlation, and coefficients close to zero indicate that our signals do not correlate.

For each of 24 languages,we randomly select $1,000$ source language words and use each of our eight basic translation signals to rank all candidate English translations. For each source language word and each pair of signals, we measure the Spearman correlation coefficient. We average the pairwise results across the $1,000$ source words and then average across languages.

Table 9 shows the results. The first thing to note is that the highest average correlation coefficient is between the frequency and the inverse-document frequency (IDF) signals (0.49). This makes sense because IDF is based on word frequency. The second highest value corresponds to a negative correlation (-0.31) between orthographic similarity and Wikipedia contextual similarity. These features are based on entirely different information, and we would not expect them to have a positive correlation. The fact that

they are negatively correlated is surprising, but confirms our intuition that the signals provide orthogonal information.

4.2 Relative Strength of Individual Signals

We analyzed the relative strength of the different signals to see if some signals tended to rank translation candidates more accurately than others. We would expect that the frequency signal is a weaker predictor than, for example, orthographic similarity, particularly for closely related language pairs. In our second set of experiments, we compare the accuracies of each signal and include analyses by language and by part-of-speech.

4.2.1 By Source Language. We computed how frequently each signal ranks the correct translation higher than any other signal. That is, we computed how often each signal is a better predictor of how to translate a given word than all other signals. We use a set of randomly selected $1,000$ source language words.[5] For each, we identify the rank of the *correct* English translation under each of the eight basic signals. We then compare how often each signal ranks the correct translation higher than the other signals. Table 10 shows the results. The following three signals dominate most often: Wikipedia contextual similarity, orthographic similarity, and topic similarity.

4.2.2 By Part of Speech. We ask a related question: are some signals particularly informative for certain classes of words? In order to begin to answer this question, we label each source word with the most probable part-of-speech (POS) tag for its English translation using the English POS tagger in the Natural Language Toolkit (Bird, Klein, and Loper 2009) to tag English words in isolation. We use information from English because POS taggers are not readily accessible for many of our languages of interest.

As before, we examine the relative performance of each signal, but breaking down the results by POS tag instead of by language. Table 11 shows the results. For clarity, we collapse some POS classes. For example, we mark both noun and plural nouns as simply 'Noun.' Because there are so few word types, we also collapse all closed class categories, including conjunctions, determiners, and prepositions into a single 'Closed' category. The final row is identical to that in Table 10. Because most (65%) words are nouns, the summary statistics are dominated by them.

The results in Table 11 are very consistent across word classes with one notable exception. The orthographic feature makes very good translation predictions for nouns and adjectives but not for the other word classes. The higher performance for orthographic similarity on nouns makes sense; we would expect orthographic similarity to be informative for borrowed and transliterated words, which tend to be proper nouns. The overall consistency suggests that there is likely little to gain from training word class-specific models for making translation predictions. In Section 4.3.1, we define a baseline method for combining the orthogonal features to make a single translation prediction, and in Section 4.3.2 we *learn* models for combining features.

5 The same randomly selected set of source words that was used in Section 4.1

Language	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.	idf
Azeri	3.6	**41.0**	3.6	11.0	30.3	5.9	4.2	0.4
Bulgarian	5.1	27.0	3.1	17.0	**42.2**	4.3	0.6	0.8
Bengali	8.7	26.7	0.9	15.4	**40.4**	4.5	2.3	1.2
Bosnian	8.8	**41.2**	4.2	16.5	21.8	4.7	2.5	0.4
Cebuano	12.7	22.1	7.3	20.6	**25.7**	4.6	6.4	0.5
Welsh	11.0	**55.6**	3.2	9.6	11.1	8.0	1.2	0.4
Gujarati	9.4	**33.9**	5.3	8.6	31.8	4.3	3.9	2.9
Hindi	4.5	25.5	2.0	10.6	**46.7**	4.9	2.8	2.9
Hungarian	4.6	**36.1**	0.0	10.1	25.7	12.5	5.4	5.7
Indonesian	12.3	**54.9**	4.3	10.8	6.4	7.9	0.5	2.8
Latvian	5.4	**41.6**	4.8	18.6	23.1	5.0	1.3	0.3
Nepali	11.2	**32.0**	6.4	12.5	27.6	5.1	4.2	0.8
Romanian	5.7	**39.3**	1.5	35.0	9.6	5.4	2.7	0.8
Slovak	4.8	**42.1**	4.2	17.5	22.8	4.3	3.3	1.0
Somali	8.7	**28.3**	3.4	11.1	18.1	17.4	12.5	0.5
Albanian	7.2	**47.8**	3.1	21.9	11.0	6.0	3.0	0.1
Serbian	3.8	27.4	1.6	17.5	**42.8**	4.5	1.6	0.7
Swedish	4.3	**45.0**	2.1	22.3	10.7	11.1	2.5	2.1
Tamil	7.7	25.2	1.8	4.2	**53.7**	5.1	1.6	0.8
Telugu	6.6	29.4	5.8	10.2	**39.9**	3.1	3.4	1.6
Turkish	6.8	**43.4**	8.7	9.8	15.2	11.4	2.5	2.1
Ukrainian	7.2	**35.1**	4.0	24.0	17.0	6.9	3.6	2.2
Uzbek	7.4	6.6	0.5	20.1	**41.0**	15.1	7.4	1.9
Vietnamese	11.0	16.6	9.7	7.7	**21.0**	16.6	3.3	14.1
Average	7.4	**34.3**	3.8	15.1	26.5	7.4	3.4	2.0

Table 10: Percent of time when each translation signal ranks a correct translation the highest out of all of the translation signals. This percentage is calculated for $1,000$ randomly chosen words with dictionary entries for each of the 24 languages.

POS Class	% Words	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.	idf
Verb	10.9	8.9	**34.0**	4.6	7.3	31.1	9.1	2.9	2.1
Noun	**64.8**	7.0	**36.7**	3.5	17.4	23.7	7.0	2.9	1.9
Adverb	3.9	10.5	**35.3**	6.6	5.1	29.0	7.3	3.5	2.6
Adjective	13.3	6.2	**34.4**	3.1	19.0	27.3	5.5	3.1	1.4
Closed	7.1	9.4	28.4	5.3	6.6	**36.8**	5.4	7.0	1.1
Average		7.4	**34.3**	3.8	15.1	26.5	7.4	3.4	2.0

Table 11: Analysis of Signals by Part-of-Speech tag. This table shows the percent of time when each translation signal ranks a correct translation highest out of all of the translation signals. The results are subdivided based on part of speech. The average row is identical to the average per-language result given in Table 10.

4.3 Accuracy of Features and their Combination

Schafer (2006) showed that combining diverse signals of translation equivalence could improve performance on bilingual lexicon induction. Here we do a more systematic analysis. We extend their observations and more systematically explore the space of possibilities by (1) experimenting with a wider variety of features, (2) analyzing a larger number of languages, and (3) introducing a discriminative model to set the weights of each feature to optimize translation quality.

4.3.1 Baseline Combination Technique: MRR. As our baseline combination, we use the mean reciprocal rank (MRR) across all monolingual signals, H,

$$MRR_e = \frac{\sum_{h \in H} \frac{1}{r_h(e)}}{|H|}$$

where $r_h(e)$ is the rank of English word e under the monolingual similarity measure h. This unsupervised approach to rank aggregation assumes no prior knowledge of which signals are likely to be the most informative.

4.3.2 Discriminative Combination of Monolingual Signals. We introduce a novel *supervised* approach to combining the monolingual signals enumerated above. For each language, we choose up to $10,000$ source language words among those that occur in each of our comparable corpora (web crawls and Wikipedia) at least ten times and that have at least one translation in our gold standard dictionaries. Because some monolingual datasets and some dictionaries are small, the source word samples are smaller than $10,000$ for some languages. For example, although our MTurk dictionary contains translations for $9,977$ Gujarati words, only $4,442$ of those words appear at least ten times in both of our monolingual corpora. We randomly divide the source language words into three equally sized sets for training, development, and testing.

We train binary classifiers to predict whether a pair of words are translations of one another or not. The translations in our training data serve as positive training examples. The negative training examples are constructed by randomly pairing source language words in the training data with English words.[6] We use our development data to set the number of negative examples per positive example. Using three negative examples for each positive example optimized performance on the development set. At test time, after scoring all source language words in the test set paired with all English words in our candidate set,[7] we rank the English candidates by their classification scores and evaluate accuracy in the top-k translations.

We use the Vowpal Wabbit package (Agarwal et al. 2014) to estimate the parameters of our classifiers. VW uses a gradient descent-based algorithm for learning binary predictors, and we perform 100 learning passes over the training data. We used the following parameters: a logistic loss function, no regularization, linear regression, and an adaptive learning rate for each feature. These choices were kept the same across all languages. Our data and software will be made available upon publication, so that other researchers may re-run our experiments and try their own models.

We train classifiers separately for each source language on a held-out development set to learn the weights of each of the 18 features. The weights vary based on, for example, corpora size and the relatedness of the source language and English (i.e. the number of cognates). Although the scale of feature values varies somewhat, making it difficult to interpret feature weights, we compared feature weights and found that the highest weighted feature for 19 languages is the Wikipedia topic similarity feature, and the highest for 5 languages is the Wikipedia context feature. These results are consistent with what we saw comparing the performance of individual features in Figure 4.

6 Among those that appear at least ten times in our monolingual data, consistent with our candidate set.

7 All English words appearing at least ten times in our monolingual data. In practice, we further limit the set to those that occur in the top-1000 ranked list according to at least one of our signals. Because words outside of these top-1000 lists are extremely unlikely to end up with a relatively high prediction score, doing so does not impact our performance but speeds up the prediction step.

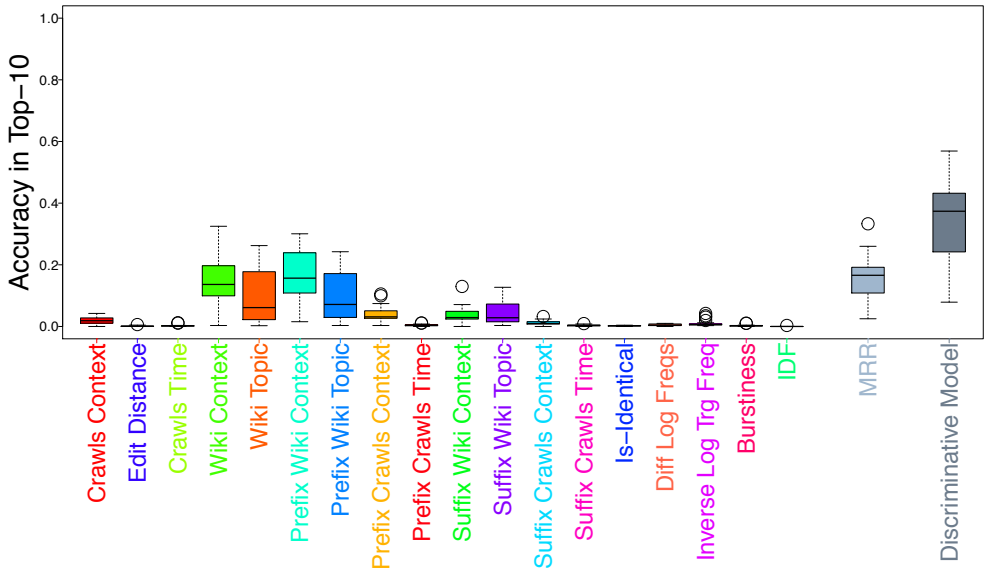


Figure 4: Performance using each of the 18 features separately to rank translation candidates, plus the MRR baseline for combining them and our discriminative model. Box and whisker plots depict the distribution of performance across a set of 24 languages. The three lines in each box illustrate the first, second (median), and third quartiles. Outliers (defined as being more than 1.5 times the interquartile range away from either quartile) are shown with circles. The whiskers show non-outlier minimum and maximum values.

4.3.3 Per-Feature Results. Figure 4 shows the performance of each of the monolingual similarity measures alone, as well as the baseline and discriminative combinations. Each box-and-whisker plot shows the top-10 accuracy range, quartiles, and median across a set of 24 diverse languages (listed in Figure 6). The Wikipedia topic and context features using whole words and word prefixes are the highest performing single features. Using the simple MRR method of combining signals is more effective than using any single feature. Our discriminative approach learn a much better way to combine the orthogonal signals, and outputs much more accurate translations.

4.3.4 Per-Language Results. For each source language, we use our trained models to induce translations for each source language word in our test sets, and we do evaluation against our gold standard bilingual dictionaries. We rank English translations by their translation classification score and measure percent accuracy in the top-k. This measure is somewhat conservative since the dictionaries aren't expected to be exhaustive, meaning that some target language translations for a given source language word won't appear in the dictionary and the system won't be given credit for ranking these target items high in its translation list. This is particularly true here because we have used the MTurk dictionaries, which are somewhat noisy. However, in these experiments, we only evaluate on words that do appear in our bilingual dictionary. It's possible that such words are easier to translate than, say, a given OOV word in some sentence which we

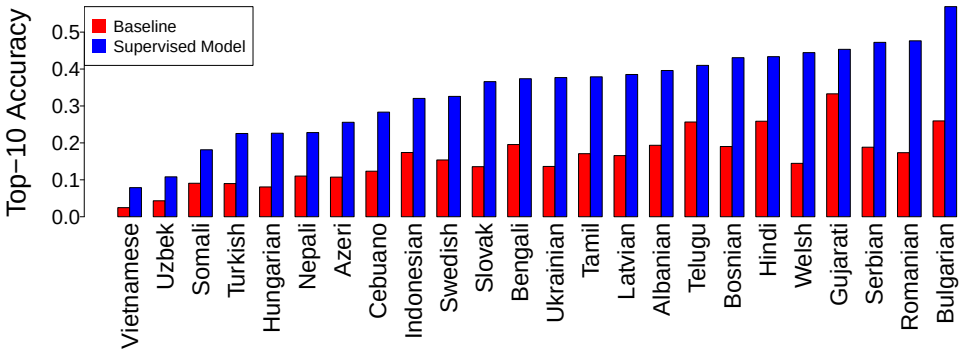


Figure 5: Top-10 bilingual lexicon induction accuracy of the baseline MRR approach to combining signals and our proposed supervised approach for each of 24 languages.

wish to translate. The results presented in this section are on the held-out blind test sets described above.

Table 12 compares the performance of the MRR baseline and our discriminative combination for each of the 24 languages. Figure 5 shows the same top-10 accuracies graphically. It's clear that the supervised method outperforms the baseline by a large margin for all 24 languages. Results using the supervised models vary from 11% accuracy on Uzbek to 57% accuracy on Bulgarian. The average accuracy across languages using the MRR baseline is 15.8% and using a supervised approach is 34.2%, or greater than *twice* the average baseline accuracy.

5. Determinants of Success

In Sections 5.1–5.3 we analyze what factors cause words to be translated accurately or inaccurately using our monolingually-derived features. We examine the amounts of monolingual and bilingual data, and the effects of word frequency and burstiness.

5.1 Learning Curve Analyses

Here we examine how accuracy changes as a function of the number of bilingual dictionary entries used to train the discriminative model, and as a function of the size of the monolingual corpora used to estimate the similarity scores that are used as features in the model.

5.1.1 Varying the Number of Translated Word Pairs. Figure 6 (at the end of the article) shows learning curves over the number of positive training instances for each source language. In all cases, the number of randomly generated negative training instances is three times the number of positive. For all languages, performance is stable after about 300 correct translations are used for training. This shows that our supervised method for combining signals requires only a small training dictionary. In most cases, for a new language, a dictionary of this size could be mined from the Internet or created using crowdsourcing (Irvine and Klementiev 2010; Pavlick et al. 2014).

Language	MRR Baseline	Supervised Model	Absolute Improvement	% Relative Improvement
Vietnamese	2.5	7.9	5.4	216.0
Uzbek	4.3	10.8	6.5	151.2
Somali	9.1	18.1	9.0	98.9
Turkish	9.0	22.5	13.5	150.0
Hungarian	8.1	22.6	14.5	179.0
Nepali	11.0	22.8	11.8	107.3
Azeri	10.7	25.6	14.9	139.3
Cebuano	12.3	28.3	16.0	130.1
Indonesian	17.4	32.0	14.6	83.9
Swedish	15.4	32.6	17.2	111.7
Slovak	13.6	36.6	23.0	169.1
Bengali	19.6	37.4	17.8	90.8
Ukrainian	13.6	37.7	24.1	177.2
Tamil	17.1	37.9	20.8	121.6
Latvian	16.6	38.5	21.9	131.9
Albanian	19.4	39.6	20.2	104.1
Telugu	25.7	41.0	15.3	59.5
Bosnian	19.0	43.1	24.1	126.8
Hindi	25.9	43.4	17.5	67.6
Welsh	14.5	44.4	29.9	206.2
Gujarati	33.3	45.3	12.0	36.0
Serbian	18.8	47.2	28.4	151.1
Romanian	17.3	47.6	30.3	175.1
Bulgarian	26.0	56.9	30.9	118.8
Average	15.8	34.2	18.3	129.7

Table 12: Top-10 Accuracy on test set. Performance increases for all languages moving from the baseline (*MRR Baseline*) to discriminative training (*Supervised Model*). The average accuracy across languages using the MRR baseline is 15.8 and using our supervised approach is 34.2.

5.1.2 Varying the Amount of Monolingual Data. How much monolingual data would we need to ensure high quality induced bilingual lexicons? Do our experiments show any signs of bilingual lexicon induction performance leveling off after a certain amount of monolingual data is available? If so, any further performance gains would have to be made by improving our underlying model, instead of taking the easier route of expanding our web crawls to additional websites. These are important considerations as we move to integrating induced translations into end-to-end SMT.

Figure 7 shows bilingual lexicon induction learning curves for four languages, Gujarati, Albanian, Azeri, and Tamil. Top 1, top 10, and top 100 accuracies are plotted on the y-axis for each language, and the x-axis shows the amount of monolingual data used to score and rank translation candidates. We generated the learning curves by sampling the web crawl and Wikipedia monolingual corpora at the same rate. The total amount of monolingual data available for Gujarati is about 5 million words, and it is about 11 million for Azeri, 13 million for Tamil, and 15 million for Albanian.

Performance levels off after about one third of the Albanian data are used. This corresponds to about 5 million words. For Gujarati, performance increases rapidly up to the full amount of 5 million monolingual words. For Tamil and Azeri, the performance continues to increase albeit at a lower rate than for Gujarati. These results indicate that we need several million words of comparable corpora to start to achieve reasonable performance, and possibly that increasing the amount of monolingual data exhibits the logarithmic improvements observed in other NLP problems like language.

5.2 Analysis by Word Frequency

Previous work on bilingual lexicon induction typically focused only on discovering translations for the most frequent words in a language. This was done for practical purposes, since the context-vector representations for high frequency words are much less sparse than for low frequency word. However, it is not a particularly realistic scenario, since for applications like SMT, the words that we would like to induce translations for are typically rare words that do not occur in our bilingual training data.

Figure 8 presents an analysis of the accuracy of our discriminative model. It bins source language words by their Wikipedia corpus frequency. We binned the words in each evaluation test set by frequency, and each bin contains 100 source language words. That is, the most frequent 100 source language words were put in the first bin, and the least frequent were put into the last bin. The x-axis in each figure plots the average corpus frequency of the words in a given bin versus the percent of those source language words that have a correct translation in the top-k ranked list of translations.

The results in Figure 8 are presented starting with the language with the least amount of Wikipedia data (Somali) and ending with the language with the largest amount (Swedish), among those languages for which results are presented. Corpus frequencies for even the most frequent words in the first few source languages are very small. For example, the average frequency of the 100 most frequent Somali words is only 13.

Prior work on bilingual lexicon induction has focused on identifying translations for frequent words. In general, our monolingual signals are stronger for those words that appear frequently in monolingual corpora than for those words that appear less frequently and have sparse context and temporal counts. Therefore, we hypothesized that translation accuracy would be higher for frequent words than for less frequent words, resulting in accuracies that go up from left to right, or from lower frequency to higher frequency, in the figures. Figure 8 shows that this effect holds true, but it is not as strong as we expected.

To quantify the effects of frequency, we compute the Spearman rank-order correlation coefficient between the frequency rank of a given source word and the rank of its correct translation.[8] Across all languages, we find a slightly positive average correlation of 0.08, indicating that, as we expected, more frequency words tend to have higher ranked correct translations. This effect is significant to a p-value of 0.01 for 14 of the 24 languages,[9] however the correlation is not as large as we expected. In the next section we conduct a similar analysis based on burstiness.

8 Although we have integer-valued frequency information, our comparison variable only contains ranks, so we convert frequency to an ordinal variable by ranking the words in each test set by their Wikipedia monolingual frequencies, from highest to lowest.

9 Bosnian, Cebuano, Somali, Nepali, Gujarati, Bengali, Latvian, Indonesian, Welsh, Tamil, Turkish, Telugu, Hungarian, Swedish

5.3 Analysis by Word Burstiness

Figure 9 presents results again on the same set of experiments but bins source language words by their Wikipedia corpus *burstiness*. We use the burstiness definition (B_w, not IDF_w) given in Section 2.6. As we did for the word frequency analysis, we bin the words in each evaluation set by burstiness, with each bin containing 100 source words. That is, the 100 most bursty source language words were put in the first bin, and the least bursty were put into the last bin. The horizontal axis in each figure plots the average burstiness of the words in a given bin versus the percent of those source language words that have a correct translation in the top-k ranked list of translations.

We hypothesized that it may be easier to induce translations for bursty words than for non-bursty words because their temporal and topic signatures are very peaked. The results in Figure 9 confirm this. Again, without binning by burstiness, we compute the Spearman rank-order correlation coefficient between the rank of a given word's burstiness and the rank of its correct translation. Across all languages, we find a positive average correlation of 0.25, indicating that, as we expected, we tend to rank correct translations higher for more bursty words. This effect is significant to a p-value of 0.01 for *all* 24 languages. Comparing our results here with those in Section 5.2, we see that burstiness is a better predictor of ranking performance on a given word than frequency.

6. Comparison with a Sophisticated Generative Model

We compare our discriminative bilingual lexicon induction approach with the popular generative model developed by Haghighi et al. (2008). Haghighi et al. (2008) presents a canonical correlation analysis (CCA) based approach to inducing bilingual lexicons. The generative model presented in that work first generates a set of one-to-one matchings, M, between pairs of source and target words. Then, a feature vector is generated for each matched word type, s_i and t_j, from a 'language-independent concept,' $z_{i,j}$. Similar to our work, source and target words are represented by feature vectors characterizing their orthographies and their contexts in monolingual corpora. However, unlike our work, the generative model proposed in Haghighi et al. (2008) allows neither source nor target word types to have multiple translations. Inference is done through boostrapped EM; the best CCA parameters, θ, are computed in the M-step, and the maximum weighted bipartite matching is found in the E-step using the Hungarian algorithm. In the first iteration, an initial lexicon is used to seed the E-step, and in additional EM iterations, an increasing number of high-confidence matchings are included until a complete bipartite matching is identified. The approach is referred to as matching canonical correlation analysis (MCCA).

Haghighi et al. (2008) presents results on three language pairs (English-Spanish, English-Chinese, and English-Arabic). However, evaluation is only done over nouns, which is a bursty word class, and lexicons are limited to high-frequency words. As we showed in Sections 5.2 and 5.3, frequent and bursty words tend to be the easiest to translate accurately.

We did the following to ensure that our comparison with MCCA is as fair as possible. We used Aria Haghighi's code to compute the translations for MCCA. We present experiments on Spanish-English, which was the best performing language pair in the MCCA paper. We use identical data sets for MCCA and our discriminative model, taking monolingual corpora from our Wikipedia collection and bilingual lexicons from our MTurk dictionary. We down-sample our data to about $6,000$ randomly selected Wikipedia page pairs (5 million words of text in both languages), to make the data

Model	Accuracy (%)
MCCA	15.1
Discriminative Model w/ Context and Orth. Features Only	24.3
Discriminative Model w/ All Features	42.3

Table 13: Comparison of bilingual lexicon induction accuracies using (1) matching canonical correlation analysis (MCCA), (2) our supervised discriminative model using only contextual and orthographic features, and (3) our supervised discriminative model using our complete feature set. Accuracy is measured as the percent of test set translations that are correctly matched by each model's full bipartite matching.

set comparable in size to Haghighi et al. (2008)'s experiments. We identify a bilingual dictionary of $1,100$ word translation pairs in the MTurk dictionary for which both the source and target lexicons are unique and all words appear in monolingual corpora greater than ten times. We use the learning parameters in Haghighi's MCCA code, which include ten iterations of bootstrapped EM and a context window of size four. We perform an experiment where our discriminative model is limited to use only the two features that the MCCA model uses (orthographic features and contextual features estimated over the Wikipedia monolingual corpora). We use MCCA to compute a full bipartite matching and measure accuracy over the complete test set of $1,000$ translation pairs.

We randomly select 100 word pairs to serve as a seed lexicon in the MCCA approach and as training data in our discriminative approach, and we use the remaining $1,000$ word pairs as an evaluation set. We use MCCA to compute a full bipartite matching and measure accuracy over the complete test set of $1,000$ translation pairs.

We use the seed lexicon of 100 word pairs to train our supervised discriminative model. As before, we randomly select three times as many negative examples for training. We then use the learned model to score all words in the source test lexicon paired with all words in the target test lexicon. In order to make our results comparable, we follow Haghighi et al. (2008) and use the Hungarian algorithm (Kuhn 1955) to find the best set of one-to-one bipartite matchings across the source and target lexicons, maximizing the total score across all matchings. We first measure the performance of our discriminative model using the orthographic and contextual features used by MCCA. Then, we also measure performance when we add our topic, frequency, and burstiness similarity features to the model.

Table 13 shows the performance of each bilingual lexicon induction model. The MCCA approach correctly matches 15% of the $1,000$ test set pairs. Our discriminative approach using only orthographic and contextual similarity features correctly matches 24%. When we add our full feature set, our model achieves 42% accuracy. These results demonstrate that our discriminative model needs no more training data than is needed to seed a generative model like the one presented in Haghighi et al. (2008). This is consistent with our results in Section 5.1.1, where we showed that our models can achieve higher accuracies on the bilingual lexicon induction task using only small amounts of supervision.

In addition to our discriminative model outperforming the MCCA generative model on the matching task, it has the added advantage of not being restricted to predicting 1:1 word translations. This is critical as, even for closely related language pairs, many words do not have a one-to-one correspondence across languages. One example

from the domain adaptation setting is the French word *enceinte*. In medical contexts, it translates as *pregnant* in English, but in government contexts it translates as *place*, *house*, or *chamber* and in scientific contexts is translates most frequently as *enclosures*. We would not want to restrict models of bilingual lexicon induction to choosing only one sense, or one translation, for French *enceinte*. That is, the polysemy of words varies across languages and it is important to be able to account for this in any model of bilingual lexicon induction.

7. Related Work

7.1 Diverse Monolingual Similarity Metrics

Schafer and Yarowsky (2002) exploit the idea that word translations tend to co-occur in time across languages, and Schafer (2006) uses this and a diverse set of other similarity measures to bootstrap a small seed bilingual dictionary and induce full dictionaries for low resource languages. Schafer (2006) combines the different signals, and weights their contribution in an ad hoc manual fashion, rather than setting them empirically by applying machine learning algorithms. Klementiev and Roth (2006) also use the temporal cue to train a phonetic similarity model for associating Named Entities across languages. Koehn and Knight (2002) use similarity in spelling as another kind of cue that a pair of words may be translations of one another. Other work has used dependency relations in place of adjacent words to define context (Garera, Callison-Burch, and Yarowsky 2009; Andrade, Matsuzaki, and Tsujii 2012).

Recent work has used graph-based models to induce translations. Mausam et al. (2010) uses freely available online dictionaries and inference over translation graphs to compile a very large, multilingual dictionary. Laws et al. (2010) use graph-based models to represent linguistic relations and induce translations. Tamura, Watanabe, and Sumita (2012) employ the classic notions of co-occurrence and contextual similarity but use graph-based label propagation to induce translations.

7.2 Other Approaches to Learning Translation of OOVs

Approaching the problem from an information retrieval perspective, Zhang, Huang, and Vogel (2005) use a system based on cross-lingual query expansion to identify translations for OOV words.

A new line of research has tried to use decipherment techniques (Knight 2013) to learn translations from monolingual corpora (Ravi and Knight 2011; Nuhn, Mauser, and Ney 2012; Dou and Knight 2012, 2013). This research line draws on previous decipherment work for solving simpler substitution/transposition ciphers, while recognizing that thinking of the foreign language as a "code" also requires customizing the decipherment algorithms so that they can deal with highly non-deterministic mappings and very large substitution tables.

7.3 Integration with machine translation

Any bilingual lexicon induction and dictionary expansion methods could be used to supplement parallel data used for estimating word alignments and scored phrase tables. The most obvious way to integrate lexicon induction output into the SMT pipeline would be to induce translations for out-of-vocabulary and rare words. That is, if a word in our test set does not have a translation in the phrase table, we could induce

one for it. Although most work on bilingual lexicon induction is motivated by the idea that outputs could be integrated into end-to-end SMT, until recently such an extrinsic evaluation was rarely performed. Daumé and Jagarlamudi (2011) use canonical correlation analysis (CCA) and both contextual and orthographic features to induce translations. Razmara et al. (2013) construct a graph using source language monolingual text and identify translations for source language OOV words by pivoting through paraphrases. In Irvine, Quirk, and Daumé (2013), we presented a method for expanding an initial translation dictionary estimated from old-domain parallel corpora by matching marginal probabilities over new-domain comparable corpora. Daumé and Jagarlamudi (2011), Razmara et al. (2013), and our prior work in Irvine, Quirk, and Daumé (2013) integrate translations into an SMT model to improve performance in domain adaptation settings.

In Klementiev et al. (2012), we described a framework for estimating the parameters of machine translation without bilingual parallel corpora. Many of the monolingually-estimated features that we used in that framework are the same as the features used here for bilingual lexicon induction. In that work, we performed oracle experiments where the translations were given by an existing phrase-table, and simply re-scored using the monolingually-estimated signals of translation equivalence.

7.4 Extracting Parallel Data from Comparable Corpora

Resnik and Smith (2003), Munteanu and Marcu (2005), Abdul-Rauf and Schwenk (2009a), Abdul-Rauf and Schwenk (2009b), and Smith, Quirk, and Toutanova (2010) identify parallel sentences in comparable corpora. Munteanu and Marcu (2006) identifies parallel sub-sentential fragments, using a probabilistic lexicon and information retrieval methods to identify similar document pairs and then uses the same word translation probabilities to detect parallel fragments within the document pairs. They supplement exiting parallel data with the new sentence and fragment pairs evaluate end-to-end SMT systems trained on the augmented parallel datasets. Quirk, Udupa, and Menezes (2007) also seek to identify phrase translation pairs from comparable corpora, but that method requires a first pass identification of promising comparable pairs of sentences from paired comparable documents. It then uses a generative model to extract fragment translation pairs. Similarly, Hewavitharana and Vogel (2011) seek to identify phrase translation pairs from comparable corpora but require a first pass to identify a set of comparable sentences and then a second pass through the data to find the best phrasal alignment within each sentence pair. These efforts at using comparable corpora to expand parallel corpora are orthogonal to the approaches that we propose in this article.

8. Conclusions

We have performed the most systematic analysis of bilingual lexicon induction to date. We analyze a set of 18 monolingually-derived signals of translation equivalence, including signals based on contextual similarity, temporal similarity, orthographic similarity, topic similarity, and features that compare the frequency and burstiness of words across languages. Analyzing the behavior of bilingual lexicon induction across two dozen languages, we find several striking conclusions.

All of the individual signals signals of translation equivalence are weak indicators by themselves. The best median performance of an individual signal reaching a mere $<$20% at ranking a translation within its top-10 prediction. The majority of signals

have ≪10% top-10 accuracy. Like Schafer and Yarowsky (2002), we find that combining diverse signals increases the translation accuracy. We can observe improvements even using a simple baseline combination method like mean reciprocal rank, although MRR performs only modestly better than the best individual signal. Our discriminative approach to combining the signals achieves dramatically improved performance. Our model outperforms the MRR baseline for all 24 languages that we experimented with, with the average top-10 accuracy more than doubling from 16% to 34%.

Although small seed dictionaries have been an essential element in bilingual lexicon induction since early work by Rapp (1995) and Fung (1995), and although much of the past research has employed multiple signals of translation equivalence, surprisingly no one has used the seed dictionary to empirically weight the contributions of the different signals.

A popular contemporary generative model, MCCA, proposed by Haghighi et al. (2008) also substantially underperforms our discriminative approach. Only a relatively small amount of bilingual data is needed to set the weights of the discriminative model. Our experiments show that having as little as 300 dictionary entries is sufficient. Moreover, we show that using a different language to set the weights for a language without a bilingual dictionary may be a successful strategy.

Our model performs well, even using relatively simple similarity estimators, like cosine distance without applying any dimensionality reduction techniques, and despite being a simple linear model. Future work could investigate additional gains from using more sophisticated models like decision trees, random forests, kernel machines or neural networks.

Additionally we present a nuanced analysis of the experiments: We quantify how diverse/orthogonal the signals of translation equivalence are by measuring the correlation of how the different signals rank the translations of 1000 words in each language. We show that the strongest individual signals (contextual similarity and topical similarity) are consistent across all languages. This is possibly due to the fact that both signals were computed using data derived from Wikipedia. This data is larger and more comparable than our other newswire data sets, and it has a higher coverage of our test words, which were themselves drawn from Wikipedia. We show that most signals are consistent across part-of-speech, except for orthographic similarity, which performs better for nouns and adjectives. We show that bilingual lexicon induction is more accurate for words that occur more frequently in monolingual corpora, and for words that exhibit more bursty behavior. We show that top-k translation accuracy can be increased by straightforwardly increasing the amount of monolingual data used to estimate the signals of translation equivalence, but that the increase appears to be log-linear or worse, requiring substantial increases in monolingual data for continued incremental gains.

Our experiments are more thorough than previous work in bilingual lexicon induction, and provide useful guidance for researchers who wish to use the techniques for applications translating out of vocabulary items for statistical machine translation. Although we focus primarily on low resource languages in this study, the techniques may also serve as potentially useful for high resource languages, which still have problems with out of vocabulary items even with there is ample bilingual training data for statistical machine translation systems.

9. 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.

Thank you to Shreejit Gangadharan for his help with refactoring code and running follow-on experiments that were suggested by the anonymous reviewers.

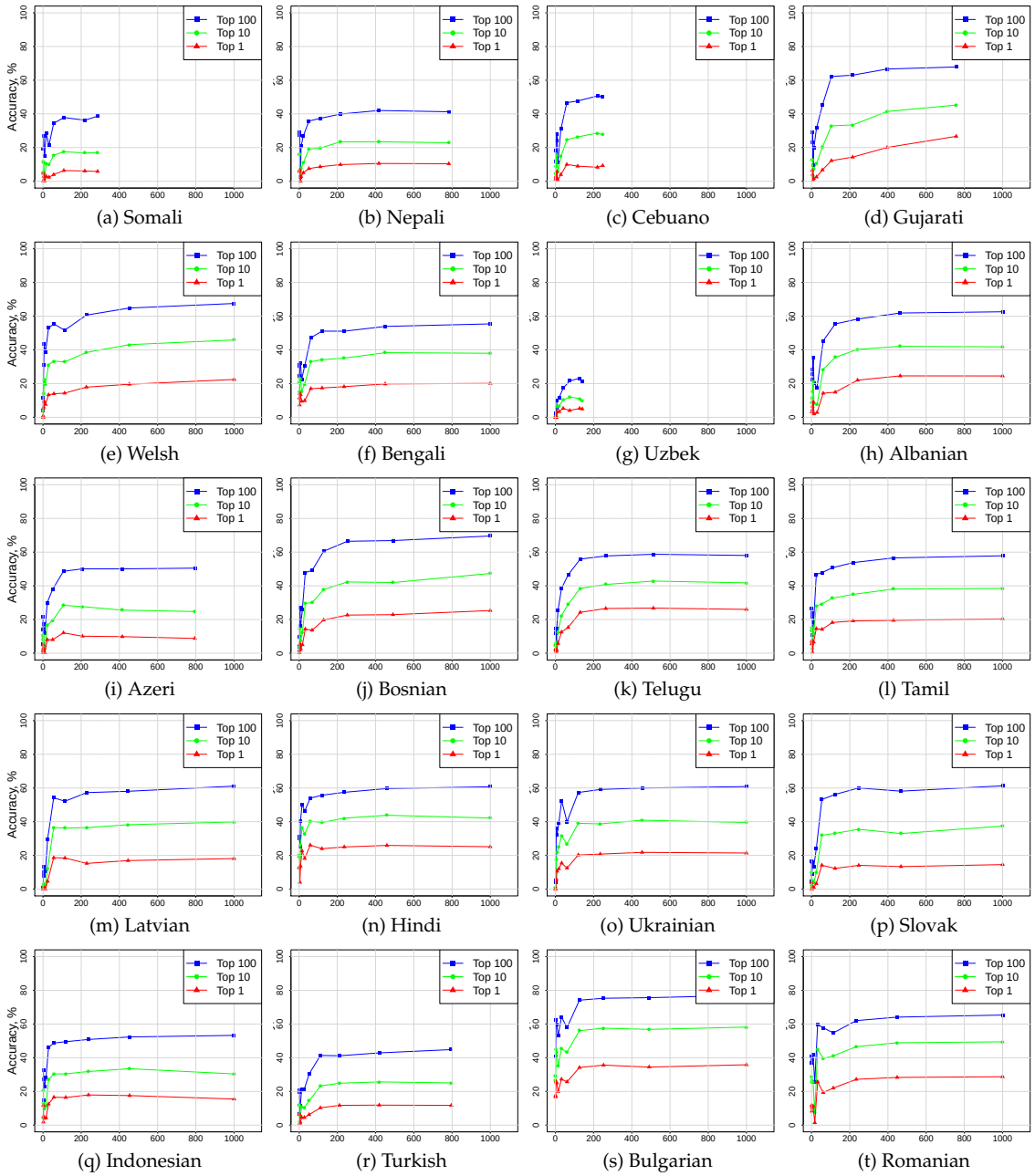
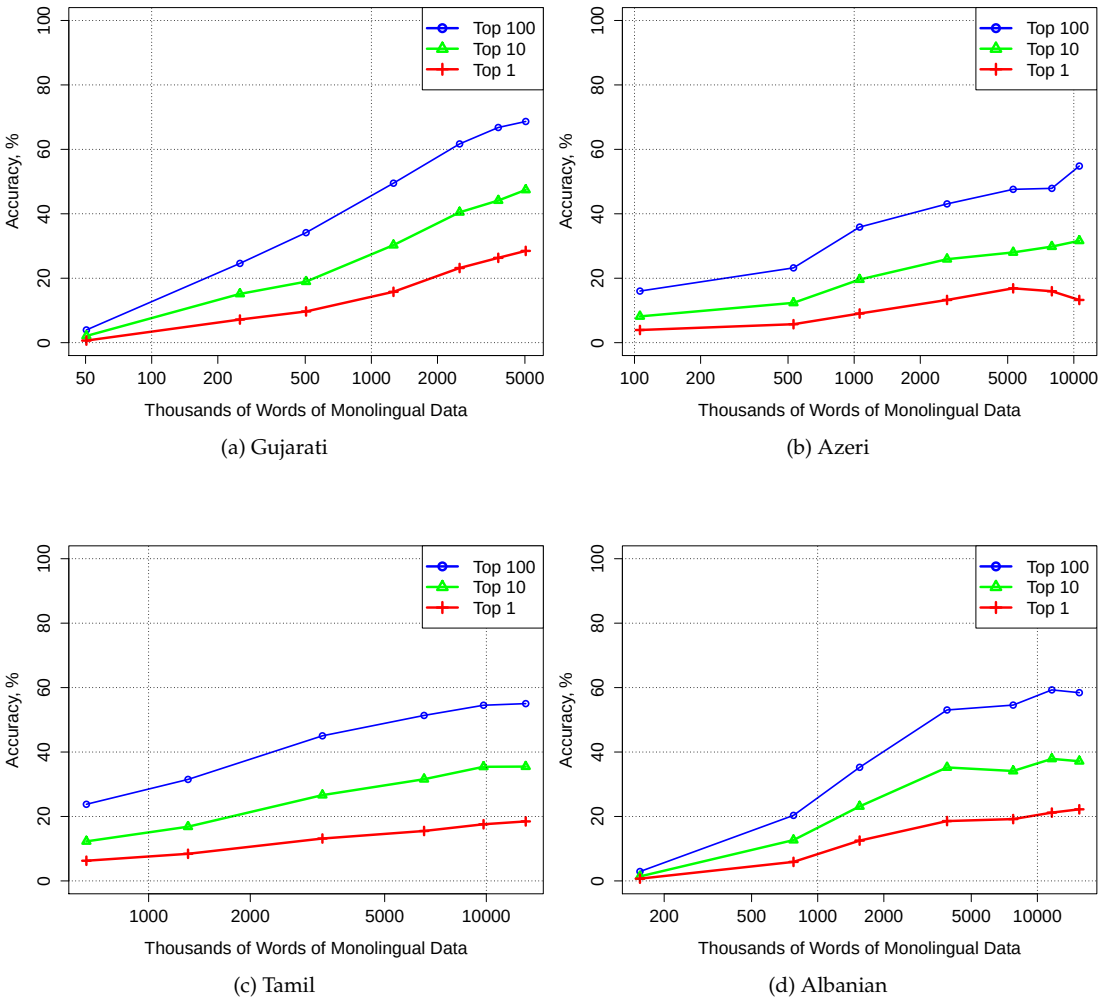


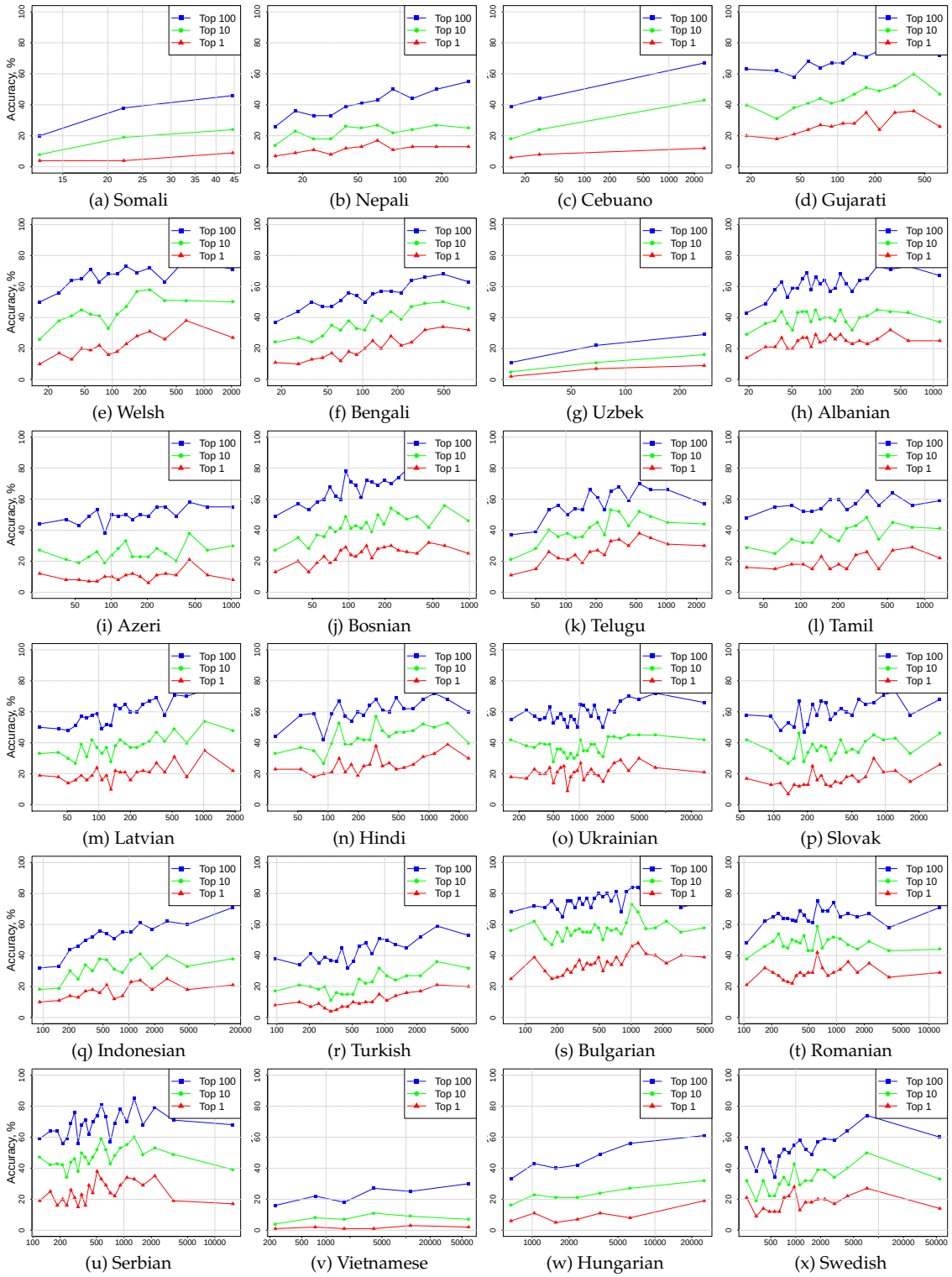
Figure 6: Learning curves over number of positive training instances, up to 1,000. For some languages, 1,000 positive training instances are not available. In all cases, the number of negative training instances is three times the number of positive. For all languages, performance is fairly stable after about 300 positive training instances.



(a) Gujarati (b) Azeri

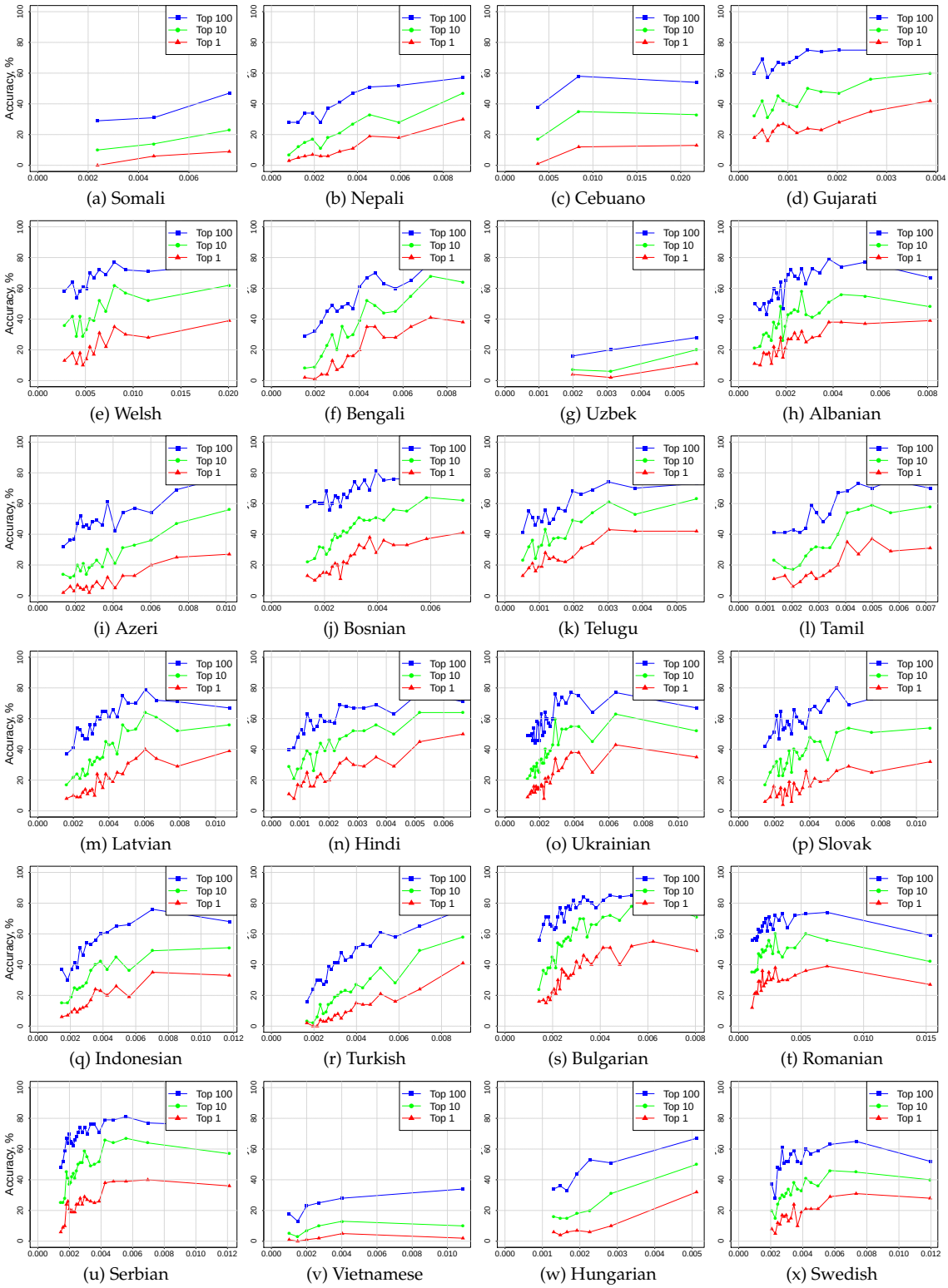
(c) Tamil (d) Albanian

Figure 7: Bilingual lexicon induction learning curves over varying comparable corpora sizes for (a) Gujarati, (b) Albanian, (c) Azeri, and (d) Tamil. The x-axis is shown on a log scale.



(a) Somali (b) Nepali (c) Cebuano (d) Gujarati
(e) Welsh (f) Bengali (g) Uzbek (h) Albanian
(i) Azeri (j) Bosnian (k) Telugu (l) Tamil
(m) Latvian (n) Hindi (o) Ukrainian (p) Slovak
(q) Indonesian (r) Turkish (s) Bulgarian (t) Romanian
(u) Serbian (v) Vietnamese (w) Hungarian (x) Swedish

Figure 8: Bilingual lexicon induction as a function of source word **frequency** in Wikipedia monolingual data. Frequency is plotted along the x-axis. Among the languages shown, we have the least monolingual data for Somali and the most for Swedish.



(a) Somali (b) Nepali (c) Cebuano (d) Gujarati
(e) Welsh (f) Bengali (g) Uzbek (h) Albanian
(i) Azeri (j) Bosnian (k) Telugu (l) Tamil
(m) Latvian (n) Hindi (o) Ukrainian (p) Slovak
(q) Indonesian (r) Turkish (s) Bulgarian (t) Romanian
(u) Serbian (v) Vietnamese (w) Hungarian (x) Swedish

Figure 9: Bilingual lexicon induction as a function of source word **burstiness** in Wikipedia monolingual data. Burstiness is plotted on the x-axis. It is calculated according to Equation 2.6.

References

Abdul-Rauf, Sadaf and Holger Schwenk. 2009a. Exploiting comparable corpora with ter and terp. In *Proceedings of the Second Workshop on Building and Using Comparable Corpora*.

Abdul-Rauf, Sadaf and Holger Schwenk. 2009b. On the use of comparable corpora to improve smt performance. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Agarwal, Alekh, Oliveier Chapelle, Miroslav Dudík, and John Langford. 2014. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133.

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)*.

Andrade, Daniel, Takuya Matsuzaki, and Jun'ichi Tsujii. 2012. Statistical extraction and comparison of pivot words for bilingual lexicon extension. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2):6:1–6:31, June.

Berg-Kirkpatrick, Taylor and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bergsma, Shane and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic, June. Association for Computational Linguistics.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

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.

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.

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. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea, July. Association for Computational Linguistics.

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.

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)*.

Garera, Nikesh, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

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

Hewavitharana, Sanjika and Stephan Vogel. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*.

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. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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)*.

Irvine, Ann and Alexandre Klementiev. 2010. Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Irvine, Ann, Chris Quirk, and Hal Daumé. 2013. Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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)*.

Knight, Kevin. 2013. Decipherment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Tutorials)*, pages 3–4, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Kuhn, Harold W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Laws, Florian, Lukas Michelbacher, Beate Dorow, Christian Scheible, Ulrich Heid, and Hinrich Schütze. 2010. A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

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.

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

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

Munteanu, Dragos and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504, December.

Nuhn, Malte, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 156–164. 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).

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

Quirk, Chris, Raghavendra Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit*.

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.

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

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380, September.

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

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)*.

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

Snyder, Benjamin, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Tamura, Akihiro, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Tao, Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 250–257, Sydney, Australia, July. Association for Computational Linguistics.

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.

Yamada, Kenji and Kevin Knight. 1999. A computational approach to deciphering unknown scripts. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Yoon, Su-Youn, Kyoung-Young Kim, and Richard Sproat. 2007. Multilingual transliteration using feature based phonetic method. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 112–119, Prague, Czech Republic, June. Association for Computational Linguistics.

Zhang, Ying, Fei Huang, and Stephan Vogel. 2005. Mining translations of OOV terms from the web through cross-lingual query expansion. In *Proceedings of the Conference on Research and Developments in Information Retrieval (SIGIR)*.