Urdu and Hindi: Translation and sharing of linguistic resources

Karthik Visweswariah, Vijil Chenthamarakshan, Nandakishore Kambhatla
IBM Research India
{v-karthik,vijil.e.c,kambhatla}@in.ibm.com

Abstract

Hindi and Urdu share a common phonology, morphology and grammar but are written in different scripts. In addition, the vocabularies have also diverged significantly especially in the written form. In this paper we show that we can get reasonable quality translations (we estimated the Translation Error rate at 18%) between the two languages even in absence of a parallel corpus. Linguistic resources such as treebanks, part of speech tagged data and parallel corpora with English are limited for both these languages. We use the translation system to share linguistic resources between the two languages. We demonstrate improvements on three tasks and show: statistical machine translation from Urdu to English is improved (0.8 in BLEU score) by using a Hindi-English parallel corpus, Hindi part of speech tagging is improved (upto 6% absolute) by using an Urdu part of speech corpus and a Hindi-English word aligner is improved by using a manually word aligned Urdu-English corpus (upto 9% absolute in F-Measure).

1 Introduction

Hindi and Urdu are official languages of India and Urdu is also the national language of Pakistan. Hindi is spoken by around 853 million people and Urdu by around 164 million people (Malik et al., 2008). Although native speakers of Hindi can comprehend most of spoken Urdu and vice versa, these languages have diverged a bit since independence of India and Pakistan – with Hindi deriving a lot of words from Sanskrit and Urdu from Persian. One clear difference between Hindi and Urdu is the script: Hindi is written in a left-to-right Devanagari script while Urdu is written in Nastaliq calligraphy style of the right-to-left Perso-Arabic script. Hence, despite the similarities, it is impossible for an Urdu speaker to read Hindi text and vice versa. The first problem we address is the translation between Hindi and Urdu in the absence of a Hindi-Urdu parallel corpus.

Though these languages together are spoken by around a billion people they are not very rich in linguistic resources. A treebank for Hindi is still under development[1] and part of speech taggers for Hindi and Urdu are trained on very small amounts of data. For translation between Hindi/Urdu and English there are no large corpora, the available corpora are an order of magnitude smaller than those available for European languages or Arabic-English. Given the lack of linguistic resources in each of the languages and the similarities between these languages, we explore whether each language can benefit from resources available in the other language.

1.1 Urdu-Hindi script conversion/translation

Sharing resources between Hindi and Urdu requires us to be able to convert from one written form to the other. Given that the languages share a good fraction of their spoken vocabularies, the obvious approach to convert between the two scripts would be to transliterate between them. While this approach has recently been attempted (Malik et al., 2009), (Malik et al., 2008) there are two main problems with this approach.

Challenges in Hindi-Urdu transliteration: Urdu uses diacritical marks that were taken from the Arabic script which serve various purposes. Urdu has short and long vowels. Short vowels are indicated by placing a diacritic with the con-

[1] https://verbs.colorado.edu/hindi_wiki/index.php/Hindi_Treebank_Data

Coling 2010: Poster Volume, pages 1283–1291,
Beijing, August 2010

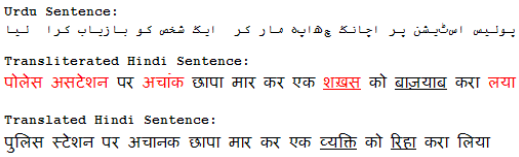

Figure 1: An Urdu sentence transliterated and translated to Hindi

sonant that precedes it in the syllable. The diacritical marks are also used for gemination (doubling of a consonant), which in Hindi is handled using a conjunct form where the consonant is essentially repeated twice. Yet another function of diacritical marks is to mark the absence of a vowel following a base consonant. Though diacritical marks are critical for correct pronunciation and sometimes even for disambiguation of certain words, they are sparingly used in written material intended for native speakers of the language. Missing diacritical marks create substantial difficulties for transliteration systems. Another difficulty is created by the fact that Urdu words cannot have a short vowel at the end of a word, whereas the corresponding Hindi word can sometimes have a short vowel. This cannot be resolved deterministically and results ambiguity in transliteration from Urdu to Hindi. A third issue is the presence of certain sounds (and their corresponding letters) that have no equivalent in Urdu. These letters are approximated in Urdu with phonetic equivalents. Transliteration from Urdu to Hindi suffers in the presence of words with these letters. Recent work on Urdu-Hindi transliteration (Malik et al., 2009) report transliteration word error rates of 16.4% and 23.1% for Urdu sentences with and without diacritical marks respectively. This problem is illustrated in Figure 1. The figure shows an Urdu sentence that is transliterated to Hindi using the Hindi Urdu Machine Transliteration (HUMT) system [2] and translated using our Statistical Machine Translation System. The words which are in red are transliteration errors (mainly because of missing diacritical marks).

Difference in Word Frequency Distributions: Even if we could transliterate perfectly between Urdu and Hindi it might not be desirable to do so from the point of view of human understanding or for machine consumption. This is because word frequencies of shared words would be different in Hindi and Urdu. At the extreme, there are several Urdu words that a fluent Hindi speaker would not understand and vice versa. More commonly, native speakers of Hindi and Urdu would use different words to refer to the same concept, even though both these words are technically correct in either of these languages. In initial experiments to quantify this issue on our corpus, which is mainly from the news domain, we estimated that around 28% of the word tokens in Urdu would not be natural in Hindi. This estimate assumes perfect transliteration, and we estimated the total error rate including transliteration at around 55% for the publicly available HUMT system. In Figure 1, the words that have been underlined have been replaced using a different word by our SMT system, even though the original word might be technically correct. Our preliminary experiments exploring this issue convinced us that to be able to convert from Urdu into natural Hindi (and vice versa) we would need to go beyond transliteration to translation to deal with the divergence of the vocabularies in the written forms of the two languages.

Importance of Context We would like to point out that in addition to word for word fidelity, there are more subtle issues in translating from Urdu-Hindi. One issue is that words in Hindi are drawn from different source languages, and with word to word translations, we might end up with phrases that are unnatural. For example, consider different ways of writing the English phrase *National* and *News* in Hindi. The word *National* in Hindi could possibly be written as *rashtriya*, *kaumi* or *national* which have origins in Sanskrit, Persian/Arabic and English respectively. Similarly the word *News* could be written as *samachar*, *khabaren* or *news* (once again with origins in Sanskrit, Persian/Arabic and English). The natural ways for writing the phrase *national news* are: *rashtriya samachar*, *kaumi khabaren* or *national news*, any of the other six combinations would be quite rare.

Another issue is that corresponding words in Hindi and Urdu might have different genders. An

[2] http://www.puran.info/HUMT/HUMT.aspx

example from (Sinha, 2009) are the words *vajah* (Urdu, feminine) and *karan* (Hindi, masculine), which would mean that the phrase *because of him* would be written as *us ke karan* in Hindi and as *us ki vajah se* in Urdu. We note that the *ke* in Hindi and *ki* in Urdu are different because of the difference in genders of the word following them. This suggests we would need to go beyond word for word translation and would need to use a higher order n-gram language model to translate with fidelity between Hindi and English.

We have established the need for going beyond transliteration, but a key challenge is to achieve good translation accuracy in the absence of a Hindi-Urdu parallel corpus. In Section 3 we describe a multi-pronged approach to translate between Hindi and Urdu in the absence of a parallel corpus that exploits the similarities between the languages.

1.2 Applications: sharing linguistic resources

We next outline the three tasks for which we consider sharing resources between Hindi and Urdu which serve as a test of the efficacy of our systems.

Statistical machine translation

In recent years, there is a lot of interest in Statistical Machine Translation (SMT) Systems (Brown et al., 1993). Modern SMT systems (Koehn et al., 2003; Ittycheriah and Roukos, 2007) learn translation models based on large amounts of parallel data. The quality of an SMT system is dependent on the amount of parallel data on which the system is trained. Unfortunately, for the pairs Urdu-English and Hindi-English, parallel data are not available in large quantities, thereby limiting the quality of these SMT systems. In this paper we show that we can improve the accuracy of an Urdu→English SMT system by using a Hindi-English parallel corpus.

Part of Speech tagging

Part of Speech (POS) tagging involves marking the part of speech of a word based on its definition and surrounding context in a sentence. Sequential modeling techniques like Hidden Markov Models (Rabiner, 1990) and Conditional Random Fields (Lafferty et al., 2001) are commonly used to build Part of Speech taggers. These models are typically trained using a manually tagged part of speech corpus. Manual tagging of data requires lot of human effort and hence large corpora are not readily available for many languages. We improve a Hindi POS tagger by using a manually tagged Urdu POS corpus.

Supervised bitext alignment

Machine generated word alignments between pairs of languages have many applications: building statistical machine translation systems, building dictionaries, projection of syntactic information to resource poor languages (Yarowsky and Ngai, 2001). Most of the early work on generating word alignments has been unsupervised, e.g. IBM Models 1-5 (Brown et al., 1993), recent improvements on the IBM Models (Moore, 2004), and the HMM algorithm described in (Vogel et al., 1996). Recently, significant improvements in performance of aligners have been achieved by the use of human annotated word alignments (Ittycheriah and Roukos, 2007; Lacoste-Julien et al., 2006). We describe a method to transfer manual word alignments from Urdu-English to Hindi-English to improve Hindi-English word alignments.

1.3 Contributions

Our main contributions are summarized below: We present a hybrid technique to translate between Hindi and Urdu in the *absence* of a Hindi-Urdu parallel corpus that significantly improves upon past efforts to convert between Hindi and Urdu via transliteration. We validate the efficacy of the translation systems we present, by using it to share linguistic resources between Hindi and Urdu for three important tasks:

1. We improve a part of speech tagger for Hindi using an Urdu part of speech corpus.

2. We use manual Urdu-English word alignments to improve the task of Hindi-English bitext alignments.

3. We use a Hindi-English parallel corpus to improve translation from Urdu to English.

2 Related work

Converting between the scripts of Hindi and Urdu is non-trivial and has been a recent focus (Malik et al., 2008; Malik et al., 2009). (Malik et al., 2008) uses hand designed rules encoded using finite state transducers to transliterate between Hindi and Urdu. As reported in (Malik et al., 2009) these hand designed rules achieve accuracies of only about 50% in the absence of diacritical marks. (Malik et al., 2009) improves Urdu→Urdu transliteration performance to 79% by post processing the output of the transducer with a statistical language model. In contrast to (Malik et al., 2009) we use a statistical model for character transliteration. As discussed in Section 1.1, due to the divergence of vocabularies in written Hindi and Urdu, transliteration is not sufficient to convert from written Urdu to written Hindi. We also use a more flexible model that allows for more natural translations by allowing Urdu words to translate into Hindi words that do not sound the same.

(Sinha, 2009) builds an English-Urdu machine translation system using an English-Hindi machine translation system and a Hindi-Urdu word mapping table, suitably adjusted for part of speech and gender. Their system is not statistical, and is largely based on manual creation of a large database of Hindi-Urdu correspondences. Additionally, as mentioned in the conclusion, their system cannot be used for direct translation from Hindi to Urdu, since a grammatical analysis of the English provides information necessary for the Hindi to Urdu mapping. In contrast to this work, our techniques are largely statistical, require minimal manual effort and can directly translate between Hindi and Urdu without the associated English.

3 Approach to translating between Hindi and Urdu

As discussed in Section 1, transliteration between Hindi and Urdu is not a straightforward task and current efforts result in fairly high error rates. We would like to combine the approaches of transliteration and translation since our goal is to use the translation for sharing linguistic resources rather than for direct consumption.

We use a fairly standard phrase based translation system to translate between Hindi and Urdu. The key challenge that we overcome is being able to develop such a system with acceptable accuracy in the absence of Hindi-Urdu resources (we have neither a parallel corpus nor a dictionary with sufficient coverage). In spite of the absence of resources, translation between this language pair is made feasible by the fact that word order is largely maintained and translation can be done maintaining a word to word correspondence. There are some exceptions to the monotonicity in the two languages. Consider the English phrase *Government of Sindh* which in Urdu would be *hukumat e sindh* in the same word order as in English, while in Hindi it would be *sindhi sarkar* with the word order flipped (with respect to English and Urdu). This example also shows that sometimes we do not have a word for word translation between Hindi and Urdu, the word *sindhi* in Hindi corresponding to the Urdu words *e sindh*. In spite of these exceptions, Hindi-Urdu translation can largely be done with the monotonicity assumption and with the assumption of word to word correspondences. Thus the central issue in translating between Hindi and Urdu is the creation of a word to word conditional probability table. We explain our technique assuming we are translating from Urdu to Hindi. We take a hybrid approach to creating this table, using three different approaches.

The first approach is the pivot language approach (Wu and Wang, 2007), with English as a pivot language. We get probabilities of a Urdu word u being generated by a Hindi word h, considering intermediate English phrases e as:

$$P_p(u|h) = \sum_e P(u|e)P(e|h)$$

The translation probabilities $P(u|e)$ and $P(e|h)$ are obtained using an Urdu-English and an English-Hindi parallel corpus respectively.

This approach works reasonably well, but suffers from a couple of drawbacks. There are several common Hindi and Urdu words for which the translation is unsatisfactory. This is because the alignments for these words are not precise, they often do not align to any English word, or align to

an English words in combination with other Hindi words. A common example of this is with verbs, consider for example the English sentence
He works
which would translate into Hindi/Urdu as:
vah kaam karta hai
with word alignments *He* ↔ *vah*, *works* ↔ *kaam karta hai* . Automatic aligners often make mistakes on these multi-word alignments, and this create problems for words like *karta* and *hai* which often do not have direct equivalents in English. To deal with this issue we manually build a small phrase table for the most frequent Hindi and Urdu words by a consulting an online Hindi-Urdu-English dictionary (Platts, 1884). We also manually handle the frequent examples we observed of cases where we need to handle differences in tokenization between Hindi and Urdu (e.g *keliye* written as one word in Urdu and as *ke liye* in Hindi).

The other issue with the pivot language approach is that for word pairs which are rare in one of the languages, $\sum_e P(u|e)P(e|h)$ can easily work out to zero. This is exacerbated by alignment errors for rarer words. Thus, to strengthen our phrase table especially for infrequent words, we use a transliteration approach to build a phrase table. Note that for rare words like names of people and places, the words in Hindi and Urdu are transliterations of each other.

In light of the issues in transliterating between Hindi and Urdu (Malik et al., 2008; Malik et al., 2009) we take a statistical approach (Abdul-Jaleel and Larkey, 2003) to building a transliteration based phrase table.

We assume a generative model for producing Urdu words from Hindi words based on a character transliteration probability table P_c. The probability $P_t(u|h)$ of generating a Urdu word u from a Hindi word h is given by:

$$P_t(u|h) = \sum_{\mathbf{a}} \prod_i P_c(u_i|h_{a(i)})P(a_i|a_{i-1}),$$

where $\mathbf{a}$ represents the alignment between the Hindi and Urdu characters, $a(i)$ is the the index of the Hindi character that the i^{th} Urdu character is aligned to, $P_c(u_c|h_c)$ is the probability of an Urdu character u_c being generated by a Hindi character h_c and $P(a_i|a_{i-1})$ represents a distortion probability. Since transliteration is monotonic and we want to encourage small jumps we set: $P(a_i|a_{i-1}) = c\eta^{(a_i - a_{i-1})}$ for $a_i > a_{i-1}$ and 0 otherwise. To obtain P_c we use the EM algorithm and we can reuse standard machinery that is used to obtain HMM word alignments in Statistical Machine Translation (with the constraint of Monotone alignments). To calculate a transliteration based phrase table, for each Hindi word h we search over a large vocabulary of Urdu words and retain words u for which $P_t(u|h)$ is sufficiently high as possible transliterations of h. We set the probabilities in the transliteration based phrase table to be proportional to $P_t(u|h)$. Finding this table requires calculating $P_t(u|h)$ for every pair of words in the Urdu and Hindi vocabulary, we use the Forward-Backward algorithm for efficiency and parallelize the calculations over several machines.

The only remaining issue is how we get training data to train our transliteration model. To obtain such training data we use a table of consonant character conversions between Hindi and Urdu as given in (Malik et al., 2008). We look for words in our pivot language based translation table, where there are at least three consonants and at least 50% of the consonants are shared. We observed that this yields pairs of words that are transliterations of one another with high precision. These word pairs are used as training data to build our character transliteration model P_c.

Final word translation table is obtained by combining our three approaches as follows: If the word is present in our dictionary, we use the translation given in the dictionary and exclude all others, if not we linearly interpolate between the probability table we get based on using English as a pivot language and probability table we get based on transliteration.

4 Experimental results

In this section we report on experiments to evaluate the quality of our translation method described in Section 3 and report on the application of Hindi↔Urdu translation to the sharing of linguistic resources between the two languages.

Algorithm 1 Create Urdu-Hindi Phrase Table

for all u such that u is very frequent Urdu word **do**
  $h \leftarrow$ Hindi word for u from dictionary
  $P_d(u|h) \leftarrow 1$
end for
$U \leftarrow$ Urdu vocabulary
$H \leftarrow$ Hindi vocabulary vocabulary
for all $u \in U, h \in H$ **do**
  $P_p(u|h) \leftarrow \sum_e P(u|e)P(e|h)$ {Create an Urdu-Hindi translation table using English as the pivot}
end for
for all $u \in U, h \in H$ such that $P_p(u|h) > \delta$ and $ConsonantOverlap(u, h) > \Delta$ **do**
  Add (u, h) to training set T
end for
$P_c \leftarrow$
$\arg\max_Q \prod_{(u,h)\in T} \sum_{\mathbf{a}} \prod_i Q(u_i|h_{a_i})P(a_i|a_{i-1})$
{Maximize using EM}
for all $u \in U, h \in H$ **do**
  $P_t(u|h) \leftarrow c \sum_{\mathbf{a}} \prod_i P_c(u_i|h_{a(i)})P(a_i|a_{i-1})$
  {Use Forward-Backward Algorithm}
end for
for all $u \in U, h \in H$ **do**
  if $P_d(u|h) \leftarrow 1$ **then**
    $P_{final}(u|h) \leftarrow 1$
  else
    $P_{final}(u|h) \leftarrow \lambda_p P_p(u|h) + \lambda_t P_t(u|h)$
  end if
end for

4.1 Evaluation of Hindi-Urdu translation

We built a Hindi-Urdu transliteration system as explained in Section 3. For building a pivot language based translation table we used 70k sentences from the NIST MT-08 corpus training corpus for Urdu-English. For Hindi-English we used an internal corpus of 230k sentences. We built our statistical transliteration model on roughly 3k word pairs that we obtained as described in Section 3. For Urdu→Hindi translation, we used a five gram language model built from a crawl of archives from Hindi news web sites (the corpus size was about 60 million words). For Hindi→Urdu translation we use the MT-08 Urdu corpus (about 1.5 million words) to build a trigram LM.

We evaluated the translation system in translating from Urdu to Hindi. We asked an annotator to evaluate 100 sentences (2700 words), by marking an error on a word if it was a wrong translation or unnatural in Hindi. We compared our translation system against the Hindi Urdu Machine Transliteration (HUMT) system[3]. We found an error rate of 18% for our system as against 46% for the HUMT system.

4.2 Word alignments

In this section we describe experiments at improving a Hindi-English word aligner using hand alignments for an Urdu-English corpus. For the Urdu-English corpus we use a manually word aligned corpus of roughly 10k sentences, while for the Hindi-English corpus we had roughly 3k sentences out of which we set aside 300 sentences (5300 words) for a test set. In addition to these (relatively) small supervised corpora we also use a sentence parallel Hindi-English corpus (without manual word alignments) of roughly 250k sentences.

For word alignments we use the Maximum Entropy aligner described in (Ittycheriah and Roukos, 2005) that is trained using hand aligned training data. We first translate the Urdu sentences in the Urdu-English word aligned corpus to Hindi, and then transfer the alignments by simply replacing the alignment links to a Urdu word by links to the corresponding decoded Hindi word. The above procedure covers bulk of the cases since Urdu-Hindi translation is largely a word to word translation. The special case of a phrase of multiple Urdu words decoded to multiple Hindi words is handled as follows: we align each of the words in the Hindi phrase to the union of the sets of English words that each word in the Urdu phrase aligns to. Once we convert the Urdu-English manual alignments to an additional corpus we build two Hindi-English alignment models, one on the original corpus, the other on the (Urdu→Hindi)-English corpus. The MaxEnt aligner (Ittycheriah and Roukos, 2005) models the probability of a

[3] http://www.puran.info/HUMT/HUMT.aspx

nTrain	Hindi data	+ Urdu
5	60.8	69.8
50	64.1	70.5
800	71.4	73.0
2800	75.1	75.7

Table 1: *Word alignment F-Measure as a function of the number of manually aligned Hindi-English sentences used for training. The third column shows improvements obtained by adding 10k Urdu-English word alignments sentences.*

num. words	$f(w_i, t_i), g(t_{i-1}, t_i)$	+ $h(t_i^u, t_i)$
5k	76.5	82.5
10k	81.7	84.7
20k	84.5	86.7
47k	90.6	91.0

Table 2: *POS tagging accuracy as a function of the amount of Hindi POS tagged data used to build the model. The third column indicates the use of the Urdu data via a feature type.*

particular set of links in the alignment L given the source sentence S and the target sentence T as: $P(L|S,T) = \prod_{i=1}^{M} p(l_i|t_1^M, s_1^K, l_1^{i-1})$. Let us denote by P_h and P_u the alignment models trained on the Hindi-English and the (Urdu→Hindi)-English corpora respectively. We combine these models log-linearly to obtain our final model for alignment:

$$P(L|S,T) = P_h^{\alpha}(L|S,T)P_u^{1-\alpha}(L|S,T).$$

To find the most likely alignment we use the same algorithm as in (Ittycheriah and Roukos, 2005) since the structure of the model is unchanged.

We report on the performance (Table 1) of a baseline Hindi-English word aligner built with varying amounts of Hindi-English manually word aligned training data compared against an aligner that combines in a model trained on the 10k (Urdu→Hindi)-English sentences. We observe large gains with small amounts of labelled Hindi-English alignment data, and even when we have 2800 sentences of Hindi-English data we see a gain in performance adding in the Urdu data. We note that the MaxEnt aligner we use (Ittycheriah and Roukos, 2005) defaults to (roughly) doing an HMM alignment using a word translation matrix obtained via unsupervised training. Thus the aligners reported on in Table 1 use a large amount of unsupervised data in addition to the small amounts of labelled data mentioned in the Table.

4.3 POS tagging

Unlike English for which there is an abundance of POS training data for Hindi and Urdu data is quite limited. For our experiments, we use the CRULP corpus (Hussain, 2008) for Urdu and a corpus from IITB (Dalal et al., 2007) for Hindi. The CRULP POS corpus has 150k words and uses a tagset of size 46 to tag the corpus. The IITB corpus has 50k words and uses a tagset of size 26. We set a side a test set of size 5k words from the IITB corpus. For part of speech tagging we use CRFs (Lafferty et al., 2001) with two types of features, $f(t_i, w_i)$ and $g(t_i, t_{i-1})$. With the small amounts of training data we have, adding additional feature templates degraded the performance.

In our POS tagging experiments we consider using the Urdu corpus to help POS tagging in Hindi. We first translate all of the CRULP Urdu data to Hindi. We cannot simply add in this data to the training data because of differences in the tagsets used in the data sets for the two languages. In order to make use of the additional Urdu POS tagged data (translated to Hindi), we build a separate POS tagger on this data, and use predictions from this model as a feature in training the Hindi POS tagger. We use these predictions via a feature template $h(t_i, t_i^u)$ where t_i^u denotes the tag assigned to the ith word by the POS tagger built from the CRULP Urdu data set translated into Hindi.

We present results in Table 2 with varying amounts of Hindi data used for training, in each case we present results with and without use of the Urdu resources. We see a small gain even when we use all of the available Hindi training data and as expected we see larger gains when smaller amounts of Hindi data are used.

We analyzed the type of errors and the error reduction when using the Urdu data for the case where we used only 5k words of Hindi data.

We find that the two frequent error types that were greatly reduced were noun being tagged as main verb (reduction of 65% relative) and main verb tagged as auxiliary verb (reduction of 71%). Reduction in confusion between nouns and main verbs is expected since these are open word classes that can most benefit from additional data. This also causes the reduction in errors of tagging main verbs as auxiliary verbs, since in Hindi, verbs are multi word groups with a main verb followed by one or more auxiliary verbs. Reduction of error rate in most of the other error types were close to the overall error rate reduction.

4.4 Sharing parallel corpora for machine translation

We experimented with using our internal Hindi-English parallel corpus (230k) sentences to obtain better translation for Urdu-English. The Urdu-English corpus we use is the NIST MT-08 training data set (70k sentences). We use the Direct Translation Model 2 (DTM) described in (Ittycheriah and Roukos, 2007) for all our translation experiments.

We build our baseline Urdu→English system using the NIST MT-08 training data. In training our DTM model we use HMM alignments, alignments with the MaxEnt aligner, and hand alignments for 10k sentences (the hand alignments were used to train the MaxEnt aligner).

We translated the Hindi in our Hindi-English corpus to Urdu, creating an additional Urdu-English corpus. We then use a MaxEnt aligner to align the Urdu-English words in this corpus. Since we expect this corpus to be relatively noisy due to incorrect translation from Urdu to Hindi we do not include this corpus while generating HMM alignments. We add the synthetic Urdu-English data with MaxEnt alignments to our baseline data and train a DTM model. Results comparing to the baseline are given Table 3, which shows an improvement of 0.8 in BLEU score over the baseline system by using data from the Hindi-English corpus.

This improvement is not due to unknown words being covered (the vocabulary covered is the same). Also note that in the bridge language approach we cannot get alternative translations for single words that were not already present in the Urdu-English phrase table. Thus, we believe that the improvement is due to longer phrases being seen more often in training. An example improved translation is shown below:

Corpus	MT08 Eval
Urdu	23.1
+Hindi	23.9

Table 3: *Improvement in Urdu-English machine translation using Hindi-English data* .

Ref: *just as long as its there they feel safe*

Baseline: *as long as this they just think there are safe*

Improved: *just as long as they are there they feel safe*

5 Conclusions

In this paper, we showed that we can translate between Hindi and English *without* a parallel corpus and improve upon previous efforts at transliterating between the two languages. We also showed that Hindi-Urdu translation can be useful to the sharing of linguistic resources between the two languages. We believe this approach to sharing linguistic resources will be of immense value especially with resources like treebanks which require a large effort to develop.

Acknowledgments

We thank Salim Roukos and Abe Ittycheriah for discussions that helped guide our efforts.

References

[AbdulJaleel and Larkey2003] AbdulJaleel, Nasreen and Leah S. Larkey. 2003. Statistical transliteration for english-arabic cross language information retrieval. In *CIKM*.

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

[Dalal et al.2007] Dalal, Aniket, Kumara Nagaraj, Uma Sawant, Sandeep Shelke, and Pushpak Bhattacharyya. 2007. Building feature rich pos tagger for morphologically rich languages. In *Proceedings of the Fifth International Conference on Natural Language Processing*, Hyderabad, India, January.

[Hussain2008] Hussain, Sarmad. 2008. Resources for urdu language processing. In *Proceedings of the 6th workshop on Asian Language Resources*.

[Ittycheriah and Roukos2005] Ittycheriah, Abraham and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *HLT/EMNLP*.

[Ittycheriah and Roukos2007] Ittycheriah, Abraham and Salim Roukos. 2007. Direct translation model 2. In Sidner, Candace L., Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 57–64. The Association for Computational Linguistics.

[Koehn et al.2003] Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

[Lacoste-Julien et al.2006] Lacoste-Julien, Simon, Benjamin Taskar, Dan Klein, and Michael I. Jordan. 2006. Word alignment via quadratic assignment. In *HLT-NAACL*.

[Lafferty et al.2001] Lafferty, J., A. McCallum, , and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

[Malik et al.2008] Malik, M. G. Abbas, Christian Boitet, and Pushpak Bhattacharyya. 2008. Hindi urdu machine transliteration using finite-state transducers. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 537–544, Manchester, UK, August. Coling 2008 Organizing Committee.

[Malik et al.2009] Malik, Abbas, Laurent Besacier, Christian Boitet, and Pushpak Bhattacharyya. 2009. A hybrid model for urdu hindi transliteration. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 177–185, Suntec, Singapore, August. Association for Computational Linguistics.

[Moore2004] Moore, Robert C. 2004. Improving ibm word alignment model 1. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 518–525, Barcelona, Spain, July.

[Platts1884] Platts, John T. 1884. *A dictionary of Urdu, classical Hindi and English*. W. H. Allen and Co.

[Rabiner1990] Rabiner, Lawrence R. 1990. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296.

[Sinha2009] Sinha, R. Mahesh K. 2009. Developing english-urdu machine translation via hindi. In *Third Workshop on Computational Approaches to Arabic-Script-based Languages*.

[Vogel et al.1996] Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA. Association for Computational Linguistics.

[Wu and Wang2007] Wu, Hua and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *ACL*.

[Yarowsky and Ngai2001] Yarowsky, David and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL*.