Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing

Matt Post[⋆] and **Chris Callison-Burch**[⋆†] and **Miles Osborne**[‡]
[⋆]Human Langage Technology Center of Excellence, Johns Hopkins University
[†]Center for Language and Speech Processing, Johns Hopkins University
[‡]School of Informatics, University of Edinburgh

Abstract

Recent work has established the efficacy of Amazon's Mechanical Turk for constructing parallel corpora for machine translation research. We apply this to building a collection of parallel corpora between English and six languages from the Indian subcontinent: Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. These languages are low-resource, under-studied, and exhibit linguistic phenomena that are difficult for machine translation. We conduct a variety of baseline experiments and analysis, and release the data to the community.

1 Introduction

The quality of statistical machine translation (MT) systems is strongly related to the amount of parallel text available for the language pairs. However, most language pairs have little or no readily available bilingual training data available. As a result, most contemporary MT research tends to opportunistically focus on language pairs with large amounts of parallel data.

A consequence of this bias is that language exhibiting certain linguistic phenomena are underrepresented, including languages with complex morphology and languages with divergent word orderings. In this paper, we describe our work gathering and refining document-level parallel corpora between English and each of six verb-final languages spoken on the Indian subcontinent: Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. This paper's contributions are as follows:

- We apply an established protocol for using Amazon's Mechanical Turk (MTurk) to collect parallel data to train and evaluate translation systems for six Indian languages.
- We investigate the relative performance of syntactic translation models over hierarchical ones, showing that syntax results in higher BLEU scores in most cases.
- We explore the impact of training data quality on the quality of the resulting model.
- We release the corpora to the research community under the Creative Commons Attribution-Sharealike 3.0 Unported License (CC BY-SA 3.0).[1]

2 Why Indian languages?

Indian languages are important objects of study for a number of reasons. These languages are low-resource languages in terms of the availability of MT systems[2] (and NLP tools in general) yet together they represent nearly half a billion native speakers (Table 1). Their speakers are well-educated, with many of them speaking English either natively or as a second language. Together with the degree of Internet penetration in India, it is reasonably straightforward to find and hire non-expert translators through crowdsourcing services like Amazon's Mechanical Turk.

[1] `joshua-decoder.org/indian-parallel-corpora`

[2] See `sampark.iiit.ac.in/sampark/web/index.php/content` for a notable growing effort.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 401–409,
Montréal, Canada, June 7-8, 2012.

செனட்டர்	அவளை	கருத்துக்கள்	தயார்
senator	*her*	*remarks*	*prepared*

Figure 1: An example of SOV word ordering in Tamil. Translation: *The senator prepared her remarks.*

হাত	কি	ল	আম
walk	*CONT*	*PAST*	*1p*

Figure 2: An example of the morphology of the Bengali word হাতছিলাম, meaning *[I] was walking*. CONT denotes the continuous aspect, while PAST denotes past tense.

In addition to a general desire to collect suitable training corpora for low-resource languages, Indian languages demonstrate a variety of linguistic phenomena that are divergent from English and understudied. One example is head-finalness, exhibited most obviously in a subject-object-verb (SOV) pattern of sentence structure, in contrast to the general SVO ordering of English sentences. One of the motivations underlying linguistically-motivated syntactic translation systems like GHKM (Galley et al., 2004; Galley et al., 2006) or SAMT (Zollmann and Venugopal, 2006) is to describe such transformations. This difference in word order has the potential to serve as a better test bed for syntax-based MT[3] compared to translating between English and European languages, most of which largely share its word order. Figure 1 contains an example of SOV reordering in Tamil.

A second important phenomenon present in these languages is a high degree of morphological complexity relative to English (Figure 2). Indian languages can be highly agglutinative, which means that words are formed by concatenating morphological affixes that convey information such as tense, person, number, gender, mood, and voice. Morphological complexity is a considerable hindrance at all stages of the MT pipeline, but particularly alignment, where inflectional variations mask patterns from alignment tools that treat words as atoms.

[3] We use *hierarchical* to denote translation grammars that use only a single nonterminal (Chiang, 2007), in contrast to *syntactic* systems, which make use of linguistic annotations (Zollmann and Venugopal, 2006; Galley et al., 2006).

language	script	family	L1
Bengali	বাংলা	Indo-Aryan	181M
Hindi	मानक हिन्दी	Indo-Aryan	180M
Malayalam	മലയാളം	Dravidian	35M
Tamil	தமிழ்	Dravidian	65M
Telugu	తెలుగు	Dravidian	69M
Urdu	اردو	Indo-Aryan	60M

Table 1: Languages. L1 is the worldwide number of native speakers according to Lewis (2009).

3 Data collection

The source of the documents for our translation task for each of the languages in Table 1 was the set of the top-100 most-viewed documents from each language's Wikipedia. These lists were obtained using page view statistics compiled from `dammit.lt/wikistats` over a one year period. We did not apply any filtering for topic or content. Table 2 contains a manually categorized list of documents for Hindi, with some minimal annotations indicating how the documents relate to those in the other languages. These documents constitute a diverse set of topics, including culture, the internet, and sex.

We collected the parallel corpora using a three-step process designed to ensure the integrity of the non-professional translations. The first step was to build a bilingual dictionary (§3.1). These dictionaries were used to bootstrap the experimental controls in the collection of four translations of each source sentence (§3.2). Finally, as a measure of data quality, we independently collect votes on the which of the four redundant translations is the best (§3.3).

3.1 Dictionaries

A key component of managing MTurk workers is to ensure that they are competently and conscientiously undertaking the tasks. As non-speakers of all of the Indian languages, we had no simple and scalable way to judge the quality of the workers' translations. Our solution was to bootstrap the process by first building bilingual dictionaries for each of the datasets. The dictionaries were then used to produce glosses of the complete source sentences, which we compared to the translations produced by the workers as a rough means of manually gauging trust (§3.2).

The dictionaries were built in a separate MTurk

PLACES	PEOPLE	PEOPLE	TECHNOLOGY	LANGUAGE AND CULTURE	RELIGION
Agra	A. P. J. Abdul Kalam	Premchand	Blog	Ayurveda	Bhagavad Gita
Bihar	Aishwarya Rai	Rabindranath Tagore	**Google**	Constitution of India	Diwali
China	Akbar	Rani Lakshmibai	Hindi Web Resources	Cricket	Hanuman
Delhi	Amitabh Bachchan	Sachin Tendulkar	**Internet**	**English language**	Hinduism
Himalayas	**Barack Obama**	Sarojini Naidu	Mobile phone	Hindi Cable News	Hinduism
India	Bhagat Singh	Subhas Chandra Bose	News aggregator	Hindi literature	Holi
Mumbai	Dainik Jagran	Surdas	RSS	Hindi-Urdu grammar	**Islam**
Nepal	Gautama Buddha	Swami Vivekananda	**Wikipedia**	Horoscope	Mahabharata
Pakistan	Harivansh Rai Bachchan	Tulsidas	YouTube	Indian cuisine	Puranas
Rajasthan	Indira Gandhi			Sanskrit	Quran
Red Fort	Jaishankar Prasad	THINGS	SEX	Standard Hindi	Ramayana
Taj Mahal	Jawaharlal Nehru	Air pollution	Anal sex		Shiva
United States	Kabir	**Earth**	**Kama Sutra**		Shiva
Uttar Pradesh	Kalpana Chawla	Essay	**Masturbation**	EVENTS	Taj Majal: Shiva Temple?
	Mahadevi Varma	Ganges	**Penis**	History of India	Vedas
	Meera	General knowledge	Sex positions	**World War II**	**Vishnu**
	Mohammed Rafi	**Global warming**	**Sexual intercourse**		
	Mohandas Karamchand Gandhi	Pollution	**Vagina**		
	Mother Teresa	Solar energy			
	Navbharat Times	Terrorism			

Table 2: The 100 most viewed Hindi Wikipedia articles (titles translated to English using inter-language links and Google translate and manually categorized). Entries in **bold** were present in the top 100 lists of at least four of the Indian top 100 lists. *Earth*, *India*, *World War II*, and *Wikipedia* were in the top 100 lists of all six languages.

language	entries	translations
Bengali	4,075	6,011
Hindi	-	-
Malayalam	41,502	144,505
Tamil	11,592	69,128
Telugu	12,193	38,532
Urdu	26,363	113,911

Table 3: Dictionary statistics. *Entries* is the number of source-language types, while *translations* lists the number of words or phrases they translated to (i.e., the number of pairs in the dictionary). Controls for Hindi were obtained using Google translate, the only one of these languages that were available at the outset of this project.

task, in which workers were asked to translate single words and short phrases from the complete set of Wikipedia documents. For each word, MTurk workers were presented with three sentences containing that word, which provided context. The control for this task was obtained from the Wikipedia article titles which are linked across languages, and can thus be assumed to be translations of each other. Workers who performed too poorly on these known translations had their work rejected.

Table 3 lists the size of the dictionaries we constructed.

3.2 Translations

With the dictionaries in hand, we moved on to translate the entire Wikipedia documents. Each human intelligence task (HIT) posted on MTurk contained ten sequential source-language sentences from a document, and asked the worker to enter a free-form translation for each. We collected four translations from different translators for each source sentence. To discourage cheating through cutting-and-pasting into automatic translation systems, sentences were presented as images. Workers were paid $0.70 per HIT. We then manually determined whether to accept or reject a worker's HITs based on a review of each worker's submissions, which included a comparison of the translations to a monotonic gloss (produced with the dictionary), the percentage of empty translations, the amount of time the worker took to complete the HIT, geographic location (self-reported and geolocated by way of the worker's IP address), and by comparing different translations of the same source segments against one another.

We obtained translations of the source-language documents in a relatively short amount of time. Figure 3 depicts the number of translations collected as a function of the amount of time from the posting of the task. Malayalam provided the highest throughput, generating half a million words in just under a

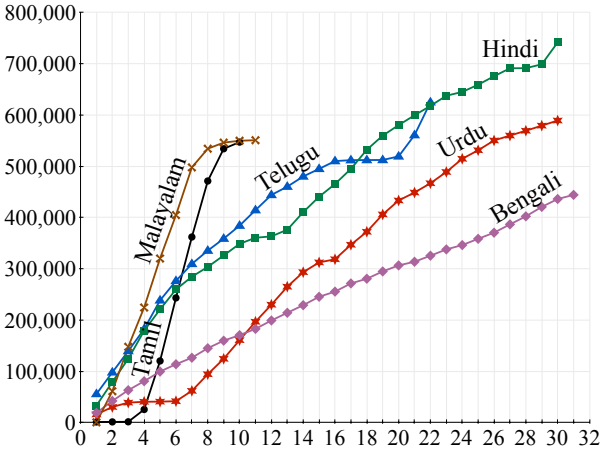

Figure 3: The total volume of translations (measured in English words) as a function of elapsed days. For Malayalam, we collected half a million words of translations in just under a week.

week. For comparison, the Europarl corpus (Koehn, 2005) has about 50 million words of English for each of the Spanish and French parallel corpora.

As has been previously reported (Zbib et al., 2012), cost is another advantage of building training data on Mechanical Turk. Germann (2001) puts the cost of professionally translated English at about $0.30 per word for translation from Tamil. Our translations were obtained for less than $0.01 per word. The rate of collection could likely be increased by raising these payments, but it is unclear whether quality would be affected by raising the base pay (although it could be improved by paying for subsequent quality control HITs, like editing).

The tradeoff for low-cost translations is increased variance in translation quality when compared to the more consistently-good professional translations. Figure 4 contains some hand-picked examples of the sorts of translations we obtained. Later, in the Experiments section (§4), we will investigate the effects this variance in translation quality has on the quality of the models that can be constructed. For now, the variance motivated the collection of an additional dataset, described in the next section.

3.3 Votes

A prevailing issue with translations collected on MTurk is the prevalence of low-quality translations. Quality suffers for a variety of reasons: Turkers lack formal training, often translate into a nonnative tongue, may give insufficient attention to the task, and likely desire to maximize their throughput (and thus their wage). Unlike Zaidan and Callison-Burch (2011), who embed controls containing source language sentences with known professional translations, we had no professionally translated data. Therefore, we could not measure the BLEU score of the Turkers.

Motivated by desire to have some measure of the relative quality and variance of the translations, we designed another task in which we presented an independent set of Turkers with an original sentence and its four translations, and asked them to vote on which was best.[4] Five independent workers voted on the translations of each source sentence. Tallying the resulting votes, we found that roughly 65% of the sentences had five votes cast on just one or two of the translations, and about 95% of the sentences had all the votes cast on one, two, or three sentences. This suggests both (1) that there was a difference in the quality of the translations, and (2) the voters were able to discern these differences, and took their task seriously enough to report them.

3.4 Data sets

For each parallel corpus, we created a standardized test set in the following manner. We first manually assigned each of the Wikipedia documents for each language into one of the following nine categories: EVENTS, LANGUAGE AND CULTURE, PEOPLE, PLACES, RELIGION, SEX, TECHNOLOGY, THINGS, or MISC. We then assigned documents to training, development, development test, and test sets in round-robin fashion using a ratio of roughly 7:1:1:1. For training data, each source sentence was repeated four times in order to allow it to be paired with each of its translations. For the development and test sets, the multiple translations served as alternate references. Table 4 lists sentence- and word-level statistics for the datasets for each language pair (these counts are prior to any tokenization).

[4] We did not collect votes for Malayalam.

மார்ச் 15,2007இல் ஆக்ஸ்ஃபோர்ட் ஆங்கில அகராதி யில் விக்கி இடம்பெற்றது.
In March 15,2007 Wiki got a place in Oxford English dictionary.
On March 15, 2007 wiki was included in the Oxford English dictionary. (5)
ON MARCH 15, 2007, WIKI FOUND A PLACE IN THE OXFORD ENGLISH DICTIONARY
March 15, 2007 oxford english index of wiki's place.

Figure 4: An example of the variance in translation quality for the human translations of a Tamil sentence; the formatting of the translations has been preserved exactly. The parenthesized number indicates the number of votes received in the voting task (§3.3).

language	dict	train	dev	devtest	test
Bengali	16k	539k	63k	61k	69k
	6k	20k	914	907	1k
Hindi	0	1,249k	67k	98k	74k
	0	37k	1k	993	1k
Malayalam	410k	664k	61k	68k	70k
	144k	29k	1k	1k	1k
Tamil	189k	747k	62k	53k	54k
	69k	35k	1k	1k	1k
Telugu	106k	951k	52k	45k	49k
	38k	43k	1k	916	1k
Urdu	253k	1,198k	67k	49k	42k
	113k	33k	736	777	605

Table 4: Data set sizes for each language pair: words in the first row, parallel sentences in the second. (The dictionaries contains short phrases in addition to words, which accounts for the difference in dictionary word and line counts.)

4 Experiments

In this section, we present experiments on the collected data sets in order to quantify their performance. The experiments aim to address the following questions:

1. How well can we translate the test sets?
2. Do linguistically motivated translation models improve translation results?
3. What is the effect of data quality on model quality?

4.1 Setup

A principal point of comparison in this paper is between Hiero grammars (Chiang, 2007) and SAMT grammars (Zollmann and Venugopal, 2006), the latter of which make use of linguistic annotations to improve nonterminal reordering. These grammars were trained with the Thrax grammar extractor using its default settings, and translated using Joshua (Weese et al., 2011). We tuned with minimum error-rate training (Och, 2003) using Z-MERT (Zaidan, 2009) and present the mean BLEU score on test data over three separate runs (Clark et al., 2011). MBR reranking (Kumar and Byrne, 2004) was applied to Joshua's 300-best (unique) output, and evaluation was conducted with case-insensitive BLEU with four references.

The training data was produced by pairing a source sentence with each of its four translations. We also added the dictionaries to the training data. We built five-gram language models from the target side of the training data using interpolated Kneser-Ney smoothing. We also experimented with a larger-scale language model built from English Gigaword, but, notably, found a drop of over a point in BLEU score. This points forward to some of the difficulties encountered with the lack of text normalization, discussed in §5.

4.2 Baseline translations

We begin by presenting BLEU scores for Hiero and SAMT translations of each of the six Indian language test sets (Table 5). For comparison purposes, we also present BLEU scores from Google translations of these languages (where available).

We observe that systems built with SAMT grammars improve measurably above the Hiero models, with the exception of Tamil and Telugu. As an external reference point, the Google baseline translation scores far surpass the results of any of our systems, but were likely constructed from much larger datasets.

Table 6 lists some manually-selected examples of

language	Hiero	SAMT	diff	Google
Bengali	12.72	13.53	+0.81	20.01
Hindi	15.53	17.29	+1.76	25.21
Malayalam	13.72	14.28	+0.56	-
Tamil	9.81	9.85	+0.04	13.51
Telugu	12.46	12.61	+0.15	16.03
Urdu	19.53	20.99	+1.46	23.09

Table 5: BLEU scores translating into English (four references). BLEU scores are the mean of three MERT runs.

the sorts of translations we obtained from our systems. While anecdotal and not characteristic of overall quality, together with the generally good BLEU scores, these examples provide a measure of the ability to obtain good translations from this dataset.

4.3 Voted training data

We noted above the high variance in the quality of the translations obtained on MTurk. For data collection efforts, there is a question of how much time and effort to invest in quality control, since it comes at the expense of simply collecting more data. We can either collect additional redundant translations (to increase quality) or translate more foreign sentences (to increase coverage).

To test this, we constructed two smaller datasets, each making use of only one of the four translations of each source sentence:

- Selected randomly
- Selected by choosing the translation that received a plurality of the votes (§3.3), breaking ties randomly (*best*)

We again included the dictionaries in the training data (where available). Table 7 contains results on the same test sets as before. These results do not clearly indicate that quality control through redundant translations are worth the extra expense. Novotney and Callison-Burch (2010) had a similar finding for crowdsourced transcriptions.

5 Further Analysis

The previous section has shown that reasonable BLEU scores can be obtained from baseline translation systems built from these corpora. While translation quality is an issue (for example, very literal translations, etc), the previous section's voted dataset experiments suggest this is not one of the most important issues to address.

இலங்கையில் சோழர் ஆட்சி
in srilanka ***solar government***
chola *rule in sri lanka*
in srilanka ***chozhas*** *ruled*
chola *reign in sri lanka*

Figure 5: An example of inconsistent orthography. Words in bold are translations of the second Tamil word.

In this section, we undertake a manual analysis of the collected datasets to inform future work. There are a number of issues that arise due to non-Roman scripts, high-variance translation quality, and the relatively small amount of training data.

5.1 Orthographic issues

Manual analysis demonstrates that inconsistencies with orthography are a serious problem. An example of this can be found in Figure 5, which contains a set of translations of a Tamil sentence. In particular, the spelling of the Tamil word சோழர் has three different realizations among the sentence's translations. The discrepancy between *zha* and *la* is due to phonetic variants (phonetic similarity may also account for the word *solar*). This discrepancy is present throughout the training and test data, where the *-la* variant is preferred to *-zha* by about 6:1 (the counts are 848 and 142, respectively).

In addition to mistakes potentially caused by foreign scripts, there are many mistakes that are simply spelling errors. Table 8 contains examples of misspellings (along with their counts) in the training portion of the Urdu-English dataset. As a point of comparison, there are no misspellings of the word in Europarl.

Such errors are present in many collections, of course, but they are particularly harmful in small datasets, and they appear to be especially prevalent in datasets like these, translated as they were by non-native speakers. Whether caused by Turker carelessness or difficulty in translation from non-Roman scripts, these are common issues, solutions for which could yield significant improvement in translation performance.

Bengali	এই সময়ই ১৯২১ সালে ঢাকা বশ্বিবদ্যিালয় স্থাপতি হয় ।
Hiero	in this time dhaka university was established on the year 1921 .
SAMT	in this time dhaka university was established in 1921 .
Malayalam	സൂര്യന്റെ ദൃശ്യമാകുന്ന ഉപരിതലത്തി ൽ താപനില 5 , 700 °k ലക്കേ് താഴ്ന്നിരിക്കും .
Hiero	the surface temperature of sun 5 , 700 degree k to down to .
SAMT	temperature in the surface of the sun 5 , 700 degree k to down to .

Table 6: Some example translations.

	Hiero		SAMT	
language	random	best	random	best
Bengali	**9.43**	9.29	**9.65**	9.50
Hindi	11.74	**12.18**	12.61	**12.69**
Tamil	**7.73**	7.48	**7.88**	7.76
Telugu	10.49	**10.61**	**10.75**	10.72
Urdu	13.51	**14.26**	14.63	**16.03**

Table 7: BLEU scores translating into English on a quarter of the training data (plus dictionary), selected in two ways: best (result of vote), and random. There is little difference, suggesting quality control may not be terribly important. We did not collect votes for Malayalam.

misspelling	**count**
japenese	91
japans	40
japenes	9
japenies	3
japeneses	3
japeneese	1
japense	1

Table 8: Misspellings of *japanese* (947) in the training portion of the Urdu-English data, along with their counts.

5.2 Alignments

Inconsistent orthography fragments the training data, exacerbating problems already present due to morpohological richness. One place this is manifested is during alignment, where different spellings mask patterns from the standard alignment techniques. We observe a large number of poor alignments, due to interactions among these problems, as well as the small size of the training data, well-documented alignment mistakes (such as garbage collecting), and the divergent sentence structures. In particular, it seems that the defacto alignment heuristics may be particularly ill-suited to these language pairs and data conditions. Figure 6 (top) contains an example of a particularly poor alignment produced by the default alignment heuristic, the *grow-diag-and* method described in Koehn et al. (2003).

As a means of testing this, we varied the alignment combination heuristics using five alternatives described in Koehn et al. (2003) and available in the `symal` program distributed with Moses (Koehn et al., 2007). Experiments on Tamil produce a range of BLEU scores between 7.45 and 10.19 (each result is the average of three MERT runs). If we plot grammar size versus BLEU score, we observe a general trend that larger grammars seem to positively correlate with BLEU score. We tested this more generally across languages using the Berkeley aligner[5] (Liang et al., 2006) instead of GIZA alignments, and found a consistent increase in BLEU score for the Hiero grammars, often putting them on par with the original SAMT results (Table 9). Manual analysis suggests that the Berkeley aligner produces fewer, more reasonable-looking alignments than the Moses heuristics (Figure 6). This suggest a fruitful approaches in revisiting assumptions underlying alignment heuristics.

6 Related Work

Crowdsourcing datasets has been found to be helpful for many tasks in natural language processing. Germann (2001) showed that humans could perform surprisingly well with very poor translations obtained from non-expert translators, in part likely because coarse-level translational adequacy is sufficient for the tasks they evaluated. That work was also pitched as a rapid resource acquisition task, meant to test our ability to quickly build systems in emergency settings. This work further demonstrates the ability to quickly acquire training data for MT systems with

[5] `code.google.com/p/berkeleyaligner/`

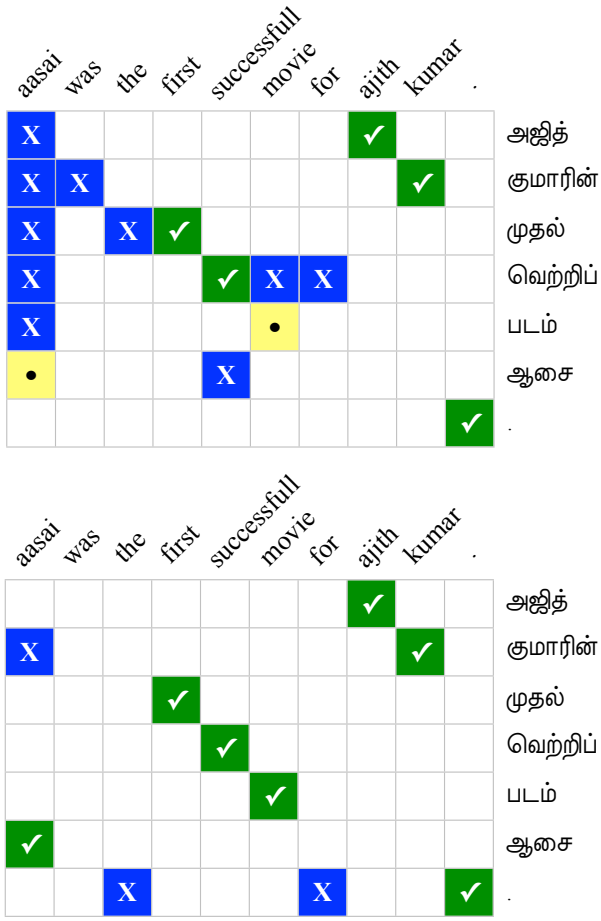

Figure 6: A bad Tamil alignment produced with the *grow-diag-and* alignment combination heuristic (top); the Berkeley aligner is better (bottom). A ✓ is a correct guess, an X marks a false positive, and a • denotes a false negative. Hiero's extraction heuristics yield 4 rules for the top alignment and 16 for the bottom.

reasonable translation accuracy.

Closely related to our work here is that of Novotney and Callison-Burch (2010), who showed that transcriptions for training speech recognition systems could be obtained from Mechanical Turk with near baseline recognition performance and at a significantly lower cost. They also showed that redundant annotation was not worthwhile, and suggested that money was better spent obtaining more data. Separately, Ambati and Vogel (2010) probed the MTurk worker pool for workers capable of translating a number of low-resource languages, including Hindi, Telugu, and Urdu, demonstrating that such workers could be found and quantifying acceptable wages and collection rates.

	grammar size			
pair	GIZA++	Berkeley	BLEU	gain
Bengali	15m	27m	13.54	**+0.82**
Hindi	34m	60m	16.47	+0.94
Malayalam	12m	27m	12.70	-1.02
Tamil	19m	30m	10.10	**+0.29**
Telugu	28m	46m	13.36	**+0.90**
Urdu	38m	58m	20.41	**+0.88**

Table 9: Hiero translation results using Berkeley alignments instead of GIZA++ heuristics. The *gain* columns denotes improvements relative to the Hiero systems in Table 5. In many cases (**bold** gains), the BLEU scores are at or above even the SAMT models from that table.

The techniques described here are similar to those described in Zaidan and Callison-Burch (2011), who showed that crowdsourcing with appropriate quality controls could be used to produce professional-level translations for Urdu-English translation. This paper extends that work by applying their techniques to a larger set of Indian languages and scaling it to training-data-set sizes.

7 Summary

We have described the collection of six parallel corpora containing four-way redundant translations of the source-language text. The Indian languages of these corpora are low-resource and understudied, and exhibit markedly different linguistic properties compared to English. We performed baseline experiments quantifying the translation performance of a number of systems, investigated the effect of data quality on model quality, and suggested a number of approaches that could improve the quality of models constructed from the datasets. The parallel corpora provide a suite of SOV languages for translation research and experiments.

Acknowledgments We thank Lexi Birch for discussions about strategies for selecting and assembling the data sets. This research was supported in part by gifts from Google and Microsoft, the EuroMatrixPlus project funded by the European Commission (7th Framework Programme), and a DARPA grant entitled "Crowdsourcing Translation". The views in this paper are the authors' alone.

References

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL*, pages 176–181. Association for Computational Linguistics.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. NAACL*, Boston, Massachusetts, USA, May.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. ACL*, Sydney, Australia, July.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: how much bang for the buck can we expect? In *ACL workshop on Data-driven methods in machine translation*, Toulouse, France, July. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. NAACL*, Edmonton, Alberta, Canada, May–June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine translation summit*.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proc. NAACL*, Boston, Massachusetts, USA, May.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *HLT-NAACL*, pages 104–111. Association for Computational Linguistics.

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proc. NAACL*, Los Angeles, California, June.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, Sapporo, Japan, July.

Jonathan Weese, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the thrax grammar extractor. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: professional quality from non-professionals. In *Proc. ACL*, Portland, Oregon, USA, June.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proc. NAACL*, Montreal, June.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, New York, New York, USA, June.