Using Mechanical Turk to Annotate Lexicons for Less Commonly Used Languages

Ann Irvine and Alexandre Klementiev
Computer Science Department
Johns Hopkins University
Baltimore, MD 21218
{anni,aklement}@jhu.edu

Abstract

In this work we present results from using Amazon's Mechanical Turk (MTurk) to annotate translation lexicons between English and a large set of less commonly used languages. We generate candidate translations for 100 English words in each of 42 foreign languages using Wikipedia and a lexicon induction framework. We evaluate the MTurk annotations by using positive and negative control candidate translations. Additionally, we evaluate the annotations by adding pairs to our seed dictionaries, providing a feedback loop into the induction system. MTurk workers are more successful in annotating some languages than others and are not evenly distributed around the world or among the world's languages. However, in general, we find that MTurk is a valuable resource for gathering cheap and simple annotations for most of the languages that we explored, and these annotations provide useful feedback in building a larger, more accurate lexicon.

1 Introduction

In this work, we make use of several free and cheap resources to create high quality lexicons for less commonly used languages. First, we take advantage of small existing dictionaries and freely available Wikipedia monolingual data to induce additional lexical translation pairs. Then, we pay Mechanical Turk workers a small amount to check and correct our system output. We can then use the updated lexicons to inform another iteration of lexicon induction, gather a second set of MTurk annotations, and so on.

Here, we provide results of one iteration of MTurk annotation. We discuss the feasibility of using MTurk for annotating translation lexicons between English and 42 less commonly used languages. Our primary goal is to enlarge and enrich the small, noisy bilingual dictionaries that we have for each language. Our secondary goal is to study the quality of annotations that we can expect to obtain for our set of low resource languages. We evaluate the annotations both alone and as feedback into our lexicon induction system.

2 Inducing Translation Candidates

Various linguistic and corpus cues are helpful for relating word translations across a pair of languages. A plethora of prior work has exploited orthographic, topic, and contextual similarity, to name a few (Rapp, 1999; Fung and Yee, 1998; Koehn and Knight, 2000; Mimno et al., 2009; Schafer and Yarowsky, 2002; Haghighi et al., 2008; Garera et al., 2008). In this work, our aim is to induce translation candidates for further MTurk annotation for a large number of language pairs with varying degrees of relatedness and resource availability. Therefore, we opt for a simple and language agnostic approach of using contextual information to score translations and discover a set of candidates for further annotation. Table 1 shows our 42 languages of interest and the number of Wikipedia articles with interlingual links to their English counterparts. The idea is that tokens which tend to appear in the context of a given type in one language should be similar to contextual tokens of its translation in the other language. Each word can thus be represented as a

Tigrinya	36	Punjabi	401
Kyrgyz	492	Somali	585
Nepali	1293	Tibetan	1358
Uighur	1814	Maltese	1896
Turkmen	3137	Kazakh	3470
Mongolian	4009	Tatar	4180
Kurdish	5059	Uzbek	5875
Kapampangan	6827	Urdu	7674
Irish	9859	Azeri	12568
Tamil	13470	Albanian	13714
Afrikaans	14315	Hindi	14824
Bangla	16026	Tagalog	17757
Latvian	22737	Bosnian	23144
Welsh	25292	Latin	31195
Basque	38594	Thai	40182
Farsi	58651	Bulgarian	68446
Serbian	71018	Indonesian	73962
Slovak	76421	Korean	84385
Turkish	86277	Ukrainan	91022
Romanian	97351	Russian	295944
Spanish	371130	Polish	438053

Table 1: Our 42 languages of interest and the number of Wikipedia pages for each that have interlanguage links with English.

vector of contextual word indices. Following Rapp (1999), we use a small seed dictionary to project[1] the contextual vector of a source word into the target language, and score its overlap with contextual vectors of candidate translations, see Figure 1. Top scoring target language words obtained in this manner are used as candidate translations for MTurk annotation. While longer lists will increase the chance of including correct translations and their morphological variants, they require more effort on the part of annotators. To strike a reasonable balance, we extracted relatively short candidate lists, but allowed MTurk users to type their own translations as well.

3 Mechanical Turk Task

Following previous work on posting NLP tasks on MTurk (Snow et al., 2008; Callison-Burch, 2009), we use the service to gather annotations for proposed bilingual lexicon entries. For 32 of our 42 languages of interest, we were able to induce lexical translation candidates and post them on MTurk for annotation. We do not have dictionaries for the remaining ten, so, for those languages, we simply posted a set of 100 English words and asked workers for manual translations. We had three distinct workers translate each word.

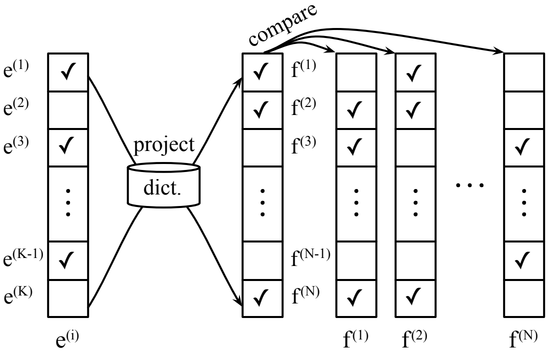

Figure 1: Lexicon induction using contextual information. First, contextual vectors are projected using small dictionaries and then they are compared with the target language candidates.

For the 32 languages for which we proposed translation candidates, we divided our set of 100 English words into sets of ten English words to be completed within a single HIT. MTurk defines HIT (Human Intelligence Task) as a self-contained unit of work that requesters can post and pay workers a small fee for completing. We requested that three MTurk workers complete each of the ten HITs for each language. For each English word within a HIT, we posted ten candidate translations in the foreign language and asked users to check the boxes beside any and all of the words that were translations of the English word. We paid workers $0.10 for completing each HIT. If our seed dictionary included an entry for a given English word, we included that in the candidate list as a positive control. Additionally, we included a random word in the foreign language as a negative control. The remaining eight or nine candidate translations were proposed by our induction system. We randomized the order in which the candidates appeared to workers and presented the words as images rather than text to discourage copying and pasting into online translation systems.

In addition to gathering annotations on candidate

[1] A simple string match is used for projection. While we expect that more sophisticated approaches (e.g. exploiting morphological analyses) are likely to help, we cannot assume that such linguistic resources are available for our languages.

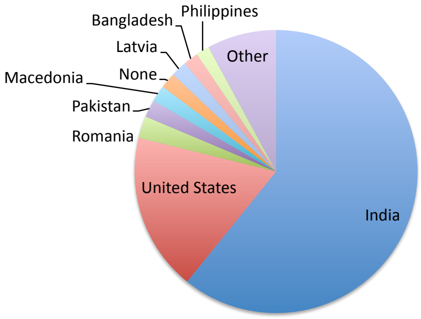

Figure 2: Distribution of MTurk workers around the world

translations, we gathered the following information in each HIT:

- Manual translations of each English word, especially for the cases where none of our proposed candidate translations were accurate
- Geographical locations via IP addresses
- How the HIT was completed: knowledge of the languages, paper dictionary, online dictionary
- Whether the workers were native speakers of each language (English and foreign), and for how many years they have spoken each

4 Results

Figure 2 shows the percent of HITs that were completed in different countries. More than 60% of HITs were completed by workers in India, more than half of which were completed in the single city of Chennai. Another 18% were completed in the United States, and roughly 2% were completed in Romania, Pakistan, Macedonia, Latvia, Bangladesh, and the Philippines. Of all annotations, 54% reported that the worker used knowledge of the two languages, while 28% and 18% reported using paper and online dictionaries, respectively, to complete the HITs.

Ninety-three MTurk workers completed at least one of our HITs, and 53 completed at least two. The average number of HITs completed per worker was 12. One worker completed HITs for 17 different languages, and nine workers completed HITs in more than three languages. Of the ten prolific workers, one was located in the United States, one in the United Kingdom, and eight in India. Because we posted each HIT three times, the minimum number of workers per language was three. Exactly three workers completed all ten HITs posted in the following languages: Kurdish, Maltese, Tatar, Kapampangan, Uzbek, and Latvian. We found that the average number of workers per language was 5.2. Ten distinct workers (identified with MTurk worker IDs) completed Tamil HITs, and nine worked on the Farsi HITs.

4.1 Completion Time

Figure 3 shows the time that it took for our HITs for 37 languages to be completed on MTurk. The HITs for the following languages were posted for a week and were never completed: Tigrinya, Uighur, Tibetan, Kyrgyz, and Kazakh. All five of the uncompleted HIT sets required typing annotations, a more time consuming task than checking translation candidates. Not surprisingly, languages with many speakers (Hindi, Spanish, and Russian) and languages spoken in and near India (Hindi, Tamil, Urdu) were completed very quickly. The languages for which we posted a manual translation only HIT are marked with a * in Figure 3. The HIT type does not seem to have affected the completion time.

4.2 Annotation Quality

Lexicon Check Agreement. Figure 4 shows the percent of positive control candidate translations that were checked by the majority of workers (at least two of three). The highest amounts of agreement with the controls were for Spanish and Polish, which indicates that those workers completed the HITs more accurately than the workers who completed, for example, the Tatar and Thai HITs. However, as already mentioned, the seed dictionaries are very noisy, so this finding may be confounded by discrepancies in the quality of our dictionaries. The noisy dictionaries also explain why agreement with the positive controls is, in general, relatively low.

We also looked at the degree to which workers agreed upon negative controls. The average percent agreement between the (majority of) workers and the negative controls over all 32 languages is only 0.21%. The highest amount of agreement with negative controls is for Kapampangan and Turkmen (1.28% and 1.26%, respectively). These are two of

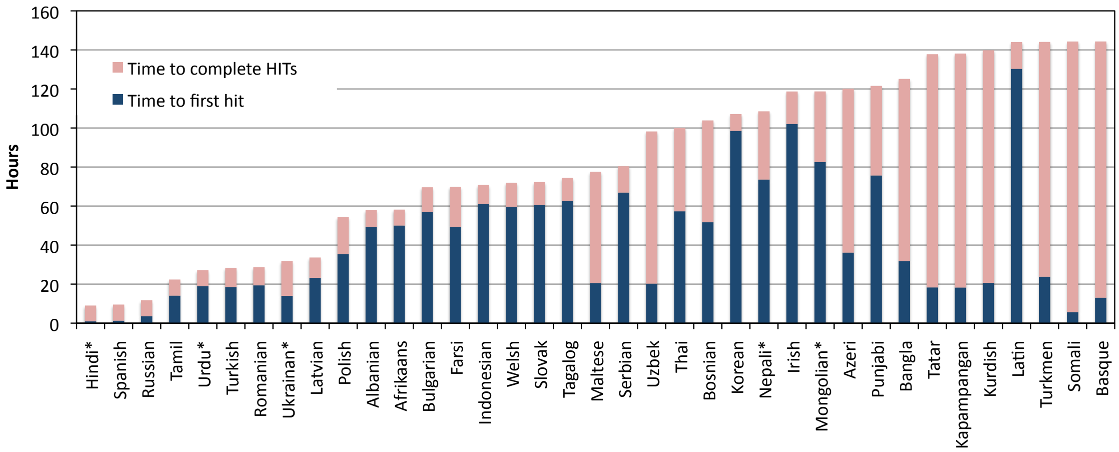

Figure 3: Number of hours HITs posted on MTurk before completion; division of the time between posting and the completion of one HIT and the time between the completion of the first and last HIT shown. HITs that required lexical translation only (not checking candidate translations) are marked with an *.

the languages for which there was little agreement with the positive controls, substantiating our claim that those HITs were completed less accurately than for other languages.

Manual Translation Agreement. For each English word, we encouraged workers to manually provide one or more translations into the foreign language. Figure 5 shows the percent of English words for which the MTurk workers provided and agreed upon at least one manual translation. We defined agreement as exact string match between at least two of three workers, which is a conservative measure, especially for morphologically rich languages. As shown, there was a large amount of agreement among the manual translations for Ukrainian, Farsi, Thai, and Korean. The MTurk workers did not provide any manual translations at all for the following languages: Somali, Kurdish, Turkmen, Uzbek, Kapampangan, and Tatar.

It's easy to speculate that, despite discouraging the use of online dictionaries and translation systems by presenting text as images, users reached this high level of agreement for manual translations by using the same online translation systems. However, we searched for 20 of the 57 English words for which the workers agreed upon a manually entered Russian translation in Google translate, and we found that the Russian translation was the top Google translation for only 11 of the 20 English words. Six of the Russian words did not appear at all in the list of translations for the given English word. Thus, we conclude that, at least for some of our languages of interest, MTurk workers did provide accurate, human-generated lexical translations.

4.3 Using MTurk Annotations in Induction

To further test the usefulness of MTurk generated bilingual lexicons, we supplemented our dictionaries for each of the 37 languages for which we gathered MTurk annotations with translation pairs that workers agreed were good (both chosen from the candidate set and manually translated). We compared seed dictionaries of size 200 with those supplemented with, on average, 69 translation pairs. We found an average relative increase in accuracy of our output candidate set (evaluated against complete available dictionaries) of 53%. This improvement is further evidence that we are able to gather high quality translations from MTurk, which can assist the lexicon induction process. Additionally, this shows that we could iteratively produce lexical translation candidates and have MTurk workers annotate them, supplementing the induction dictionaries over many iterations. This framework would allow us to gener-

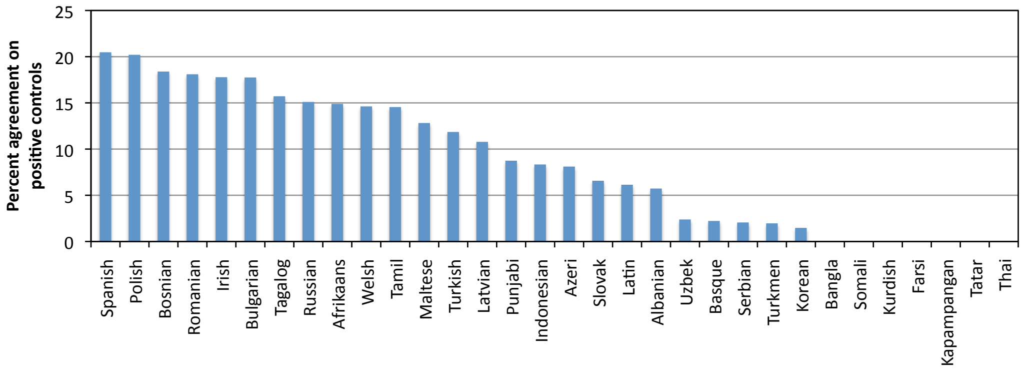

Figure 4: Percent of positive control candidate translations for which two or three workers checked as accurate.

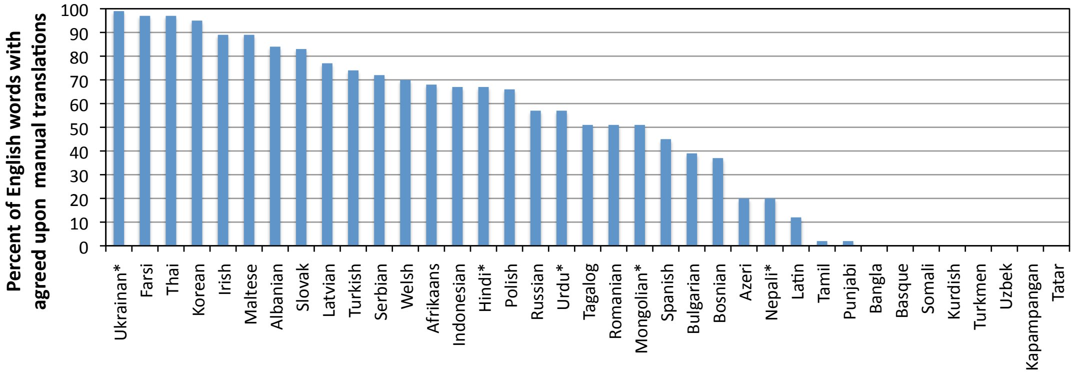

Figure 5: Percent of 100 English words for which at least two of three MTurk workers provided at least one matching manual translation; HITs that required lexical translation only (not checking candidate translations) are marked with an *.

ate very large and high quality dictionaries starting with a very small set of seed translation pairs.

5 Conclusion

The goal of this work was to use Amazon's Mechanical Turk to collect and evaluate the quality of translation lexicons for a large set of low resource languages. In order to make the annotation task easier and maximize the amount of annotation given our budget and time constraints, we used contextual similarity along with small bilingual dictionaries to extract a set of translation candidates for MTurk annotation. For ten of our languages without dictionaries, we asked workers to type translations directly. We were able to get complete annotations of both types quickly for 37 of our languages. The other five languages required annotations of the latter type, which may explain why they remained unfinished.

We used annotator agreement with positive and negative controls to assess the quality of generated lexicons and provide an indication of the relative difficulty of obtaining high quality annotations for each language. Not surprisingly, annotation agreement tends to be low for those languages which are especially low resource, as measured by the number of Wikipedia pages. Because there are relatively few native speakers of these languages in the online community, those HITs were likely completed by non-native speakers. Finally, we demonstrated that augmenting small seed dictionaries with the obtained lexicons substantially impacts contextual lexicon induction with an average relative gain of 53% in accuracy across languages.

In sum, we found that the iterative approach of automatically generating noisy annotation and asking MTurk users to correct it to be an effective means of obtaining supervision. Our manual annotation tasks are simple and annotation can be obtained quickly for a large number of low resource languages.

References

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazons mechanical turk. In *Proceedings of EMNLP*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of ACL*, pages 414–420.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2008. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of CoNLL*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL*, pages 771–779.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of AAAI*.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*, pages 519–526.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of CoNLL*, pages 146–152.

Rion Snow, Brendan OConnor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.