Translation Induction on Indian Language Corpora using Translingual Themes from Other Languages

Goutham Tholpadi and Chiranjib Bhattacharyya and Shirish Shevade

Computer Science and Automation
Indian Institute of Science
Bangalore 560012 India
{gtholpadi,chiru,shirish}@csa.iisc.ernet.in

Abstract. Identifying translations from comparable corpora is a well-known problem with several applications, e.g. dictionary creation in resource-scarce languages. Scarcity of high quality corpora, especially in Indian languages, makes this problem hard, e.g. state-of-the-art techniques achieve a mean reciprocal rank (MRR) of 0.66 for English-Italian, and a mere 0.187 for Telugu-Kannada. There exist comparable corpora in many Indian languages with other "auxiliary" languages. We observe that translations have many topically related words in common in the auxiliary language. To model this, we define the notion of a *translingual theme*, a set of topically related words from auxiliary language corpora, and present a probabilistic framework for translation induction. Extensive experiments on 35 comparable corpora using English and French as auxiliary languages show that this approach can yield dramatic improvements in performance (e.g. MRR improves by 124% to 0.419 for Telugu-Kannada). A user study on *WikiTSu*, a system for cross-lingual Wikipedia title suggestion that uses our approach, shows a 20% improvement in the quality of titles suggested.

1 Introduction

The task of identifying translations for terms is usually posed as one of generating translation correspondences. A translation correspondence for a source word assigns a score to every target word proportional to its topical similarity to the source word, so that the translation is assigned the highest score. Translation correspondences are key inputs for building human readable dictionaries, as well as for many language processing systems, including machine translation and cross language information retrieval [1].

Comparable corpora-based[1] translation correspondence induction (CC-TCI) is a popular approach for obtaining translation correspondences. Most methods using this approach require dictionaries and parsers, or make assumptions about

[1] "Comparable corpora" are document-aligned multilingual corpora, where the aligned documents are in different languages and "talk about the same thing" [2].

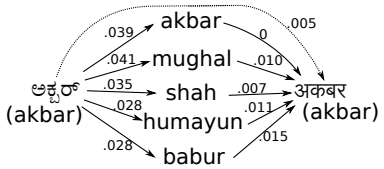

Fig. 1. A subset of the *translingual theme* in English (words in center) for a Kannada (left)–Marathi (right) translation pair. The arrow from w_1 to w_2 is labeled with the probability $P_{\mathrm{CC}}(w_2|w_1)$ (see Section 3.2)

properties of the languages involved (see Section 2). However, for many language pairs such as in Indian languages, the CC-TCI problem poses several challenges:

- Resources such as seed bilingual lexicons and linguistic tools (POS taggers, morpho-syntactic analyzers, etc.) required by some methods (e.g. [3], [4]) are not be available.
- Language properties such as presence of cognates, and orthographic similarity, cannot be assumed in general, ruling out some methods (e.g. [5], [6]).
- The only available cross-language resource is a comparable corpus. However, even this is relatively small for most language pairs, so that "CC-only" methods (e.g. [7], [8]) do not perform well.

We observe that source and target translations have many topically related words in common in other "auxiliary" language corpora[2], which can be a useful cue for identifying translations. To model this, we define the notion of a *translingual theme* (for a source–target word pair) as a set of words derived from auxiliary language comparable corpora that statistically co-occur with the source and target words. For example, Figure 1 shows the source–target pair ಅಕ್ಬರ್ /akbar/ and अकबर /akbar/ (both referring to the proper noun "Akbar"[3]) from a Kannada–Marathi corpus, and a subset {'mughal', 'shah', 'humayun', 'babur'}[4] of its translingual theme derived from Kannada–English and Marathi–English auxiliary corpora. In this work, we investigate the utility of *auxiliary* language corpora for boosting CC-TCI performance. For this purpose, we leverage Wikipedia, a large web-based multilingual encyclopedia with more than 26 million articles in 285 languages. In Wikipedia, articles in different languages on the same topic are linked (by "`langlink`"s), which enables us to quickly construct corpora for a large number of language pairs.

Cross-lingual Wikipedia Title Suggestion. The proportion of content in Wikipedia in different languages varies widely [9], and the topics covered also

[2] Comparable corpora where one language is from the pair under consideration, and the other can be any other (auxiliary) language.

[3] Akbar was a king from the Mughal dynasty who ruled parts of North India in the 16^{th} century A.D.

[4] Shah is a royal title; Humayun and Babur were both Mughal kings.

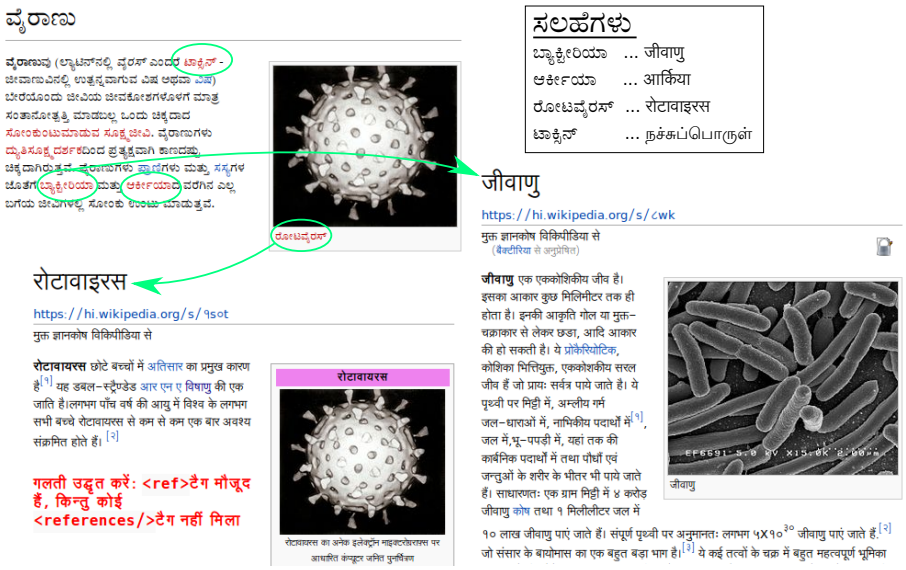

Fig. 2. A multilingual user reading a Kannada article on ವೈರಾಣು ("virus") (top-left) finds the words ಟಾಕ್ಸಿನ್ ("toxin"), ಬ್ಯಾಕ್ಟೀರಿಯಾ ("bacteria"), ಆರ್ಕೀಯಾ ("Archaea") and रोटावाइरस ("rotavirus") interesting, but there are no Kannada articles for these concepts. In response, the system gives Wikipedia title suggestions (box at top-right) from Hindi and Tamil (जीवाणु ("bacteria"), and so on).

vary with language. If a Wikipedia concept has no article in one language, articles in other languages might be suggested to a multilingual user. For example (see Figure 2), an Indian user browsing the Kannada article ವೈರಾಣು /ʋajraːɳʊ/ ('virus') might want to know about ಬ್ಯಾಕ್ಟೀರಿಯಾ /bjaːkʈiːrijaː/ ('bacteria'), and ರೋಟವೈರಸ್ /roːʈaʋajras/ ('rotavirus'). There are no articles for these concepts in Kannada, but there *are* articles in Hindi, viz. जीवाणु /dʒiːʋaːɳu/ ('bacteria') and रोटावाइरस /roːʈaːʋaːiras/ ('rotavirus'). These titles can be suggested to the user (the box at top-right in the figure) for further reading. Recently, [10] attempted a similar task using `langlinks`, where the setting was restricted to source words that are Wikipedia titles. The task of suggesting target-language Wikipedia titles for *source words that are not Wikipedia articles* has not been attempted before. In the absence of `langlinks`, this task is difficult to solve, especially for under-resourced languages without machine translation (MT), dictionaries, parsers, and parallel corpora. In this resource-scarce setting, we attempted the title suggestion task using a CC-TCI approach, leveraging auxiliary language corpora from Wikipedia. The resulting system *WikiTSu* can work for any Wikipedia language pair, and uses a Wikipedia corpus as the *only* resource.

Contributions. Our main contributions are:
• We define a new probabilistic notion of cross-language similarity in the context of comparable corpora. We show how this notion naturally admits auxiliary language corpora under certain assumptions. We also show how to combine similarities from multiple auxiliary languages using a simple mixture model, and use the combined score for translation correspondence induction. (Section 3.1)
• We perform extensive experiments on 35 comparable corpora in 9 languages from 4 language families (Indo-Aryan, Dravidian, Germanic, Romance) extracted from Wikipedia, and show significant boosts (upto 124%) in performance for a state-of-the-art CC-TCI method. (Section 4.2)
• To address the cross-lingual Wikipedia title suggestion task for the difficult resource-scarce setting, we built a system *WikiTSu* that works for *any* language pair in Wikipedia, using *no other resources*. We show via a user study that *WikiTSu* does significantly better than a state-of-the-art baseline. (Section 4.4)
• We are releasing translation correspondences for 42 language pairs (nearly 5000 words per language, 10 candidates per word) for public use as probabilistic dictionaries, or as inputs to annotator tools for dictionary building. As of today, there exist *no* dictionaries for most of these language pairs.
• We are making publicly available[5] a large curated collection of comparable corpora and gold standard translation pair sets in 7 under-resourced languages. We are also releasing the code for *WiCCX*, an in-house tool for generating preprocessed and algorithm-ready comparable corpora from Wikipedia dumps.

2 Related Work

Translation Correspondence Induction using Comparable Corpora. The problem of inducing translation correspondences from bilingual comparable corpora was introduced by [11]. There have been several approaches to this task, differentiated by the resource assumptions made.

Knowledge-based Approaches. Many approaches to translation correspondence induction use seed lexicons [2][4][12,13,14,15], syntactic/morphological analyzers [16,17,18,19], parallel corpora, translation/transliteration models [20], and other resources [3,21]. Other approaches make assumptions about the languages or corpora, such as syntactic structure, orthographic similarities, presence of cognates, monogenetic relationships, domain-specific content [5,6][22,23,24,25,26]. [27] and [28] use existing dictionaries to induce translation correspondences. There is also work on comparable corpora-based named entity mining [29,30,31] which has a similar setting, but addresses a different problem. [9] use canonical correlation analysis for Wikipedia name search, and [32] use Wikipedia link structure for translation correspondence induction. These are complementary to our statistical approach, and they can be combined to improve performance.

[5] `http://www.cicling.org/2015/data/31`

Comparable Corpora-only Approaches. [7] and [33] proposed methods that use only comparable corpora and were applied to relatively high quality corpora. The most recent work using only comparable corpora is by [8] and [34] who use latent space models, and demonstrate good performance on Wikipedia data.

Improving CC-TCI. There have been efforts to improve the results from existing methods by pre- or post-processing. [35] and [36] attempt to improve corpus quality before doing translation correspondence induction. [37] take a noisy translation correspondence obtained from any method and incorporates knowledge from *monolingual corpora in the languages of the pair* to improve accuracy. Our method, on the other hand, takes a noisy translation correspondence and incorporates knowledge from *comparable corpora in auxiliary languages* to improve accuracy. These approaches are complementary to our approach, and they can be combined to improve accuracy further.

Combination Approaches. [16] represent different kinds of relationships between words on a graph and use SimRank [38] to compute a combined score. [21] combine information with a mixture model similar to ours, while [25] use a voting scheme instead.

Using Auxiliary Languages. [39] attempted to use auxiliary languages for translation correspondence induction, but using parallel corpora. [40], [1], [27], and [41] use existing dictionaries or monogenetic relationships, while we work in the comparable corpora-only setting and make no assumptions about the language family. Auxiliary language approaches have also been used for other problems, e.g. *triangulation* for machine translation [42,43,44], word alignment [45], transliteration [46], and paraphrase extraction [47].

3 Problem Formulation and Approach

3.1 Problem Definition

Let L_S and L_T denote the source and target languages, with vocabularies V_S and V_T respectively. The translation correspondence for $s \in V_S$ is the set $TC(s) = \{(t, r_{st})\}_{t \in V_T}$ where $r_{st} \in [0, \infty)$ is the topical similarity of t to s. A translation correspondence can be viewed as being generated from a scoring function $S_{raw}()$ such that $S_{raw}(t|s) = r_{st}$. Given a comparable corpus, any method in Section 2 can be used to learn the scoring function $S_{raw}(t|s)$.[6] This function induces a ranking over the words in V_T for each word s in V_S. We assume that there exists an auxiliary language L_A which has comparable corpora with L_S and L_T, so that we can learn scoring functions $S_{raw}(a|s)$, $S_{raw}(s|a)$, $S_{raw}(t|a)$ and $S_{raw}(a|t)$, analogous to $S_{raw}(t|s)$.

The objective is to compute a scoring function $S_A(t|s)$ that uses the S_{raw} scoring functions and gives a better ranking over V_T for each s.

[6] We use the method by [8] to obtain S_{raw}, and also as the baseline (Section 4.1).

3.2 Incorporating Information From an Auxiliary Language

Cross-language Similarity in Terms of a Comparable Corpus. A document-aligned multilingual comparable corpus in l languages can be viewed as a set of tuples (each tuple contains l documents, one per language). Consider a random experiment where we sample a word from one of the documents of such a tuple. Define the random variables: $S \triangleq$ the word sampled from the L_S–document in the tuple; $T \triangleq$ the word sampled from the L_T–document in the tuple. Let $P_{\text{CC}}(T = t|S = s)$ be the probability that the sampled L_T–word is t given that a sampled L_S–word is s. This probability will be high for values of t (i.e. L_T–words) that are topically related to s. For example, given that we sampled ಬ್ಯಾಕ್ಟೀರಿಯಾ /bjaːkʈiːrijaː/ ('bacteria') from the L_S–document, we are very likely to sample words like जीवाणु /dʒiːvaːɳu/ ('bacteria') or रोग /roːg/ ('disease') from the L_T–document.[7] This is similar in spirit to the idea of *lexical triggers* [48]. We can use a baseline scoring function S_{raw} (as defined in Section 3.1) and define the *trigger probability* $P_{\text{CC}}(t|s) \triangleq \frac{S_{raw}(t,s)}{\sum_{t'} S_{raw}(t',s)}$.[8] This models topical relatedness in the context of comparable corpora in a probabilistic setting.[9] Since this model is asymmetric, i.e. in general $P_{\text{CC}}(t|s) \neq P_{\text{CC}}(s|t)$, we can expect that the translation induction performance depends on the choice of the source language, and this is confirmed by our experiments (Section 4.2).

Translingual Themes. Define the random variable $A \triangleq$ the word sampled from the L_A document in the tuple. Similar to $P_{\text{CC}}(t|s)$, we get $P_{\text{CC}}(t|a)$ and $P_{\text{CC}}(a|s), \forall a \in V_A$. We define the *source theme* for s as the set $ST_A(s) \subset V_A$ that satisfies $\forall a \in ST_A(s), a' \in V_A \setminus ST_A(s), P_{\text{CC}}(a|s) \geq P_{\text{CC}}(a'|s)$, and $\sum_{a \in ST_A(s)} P_{\text{CC}}(a|s) < \tau$, where $\tau < 1$ is threshold determined empirically. The source theme is a set of L_A words that have the highest trigger probability given the source word s. We define the *target theme* for t as the set $TT_A(t) = \{a|t \in ST_T(a)\}$, i.e. the target theme is the set of L_A words for which the target word t has a high trigger probability. Finally, we define the *translingual theme* for the ordered pair (t, s) as $TLT_A(t, s) = ST_A(s) \cap TT_A(t)$.

Using Translingual Themes to Compute Word Similarity. Our probabilistic definition allows us to write $P_{\text{CC}}(t|s) = \sum_{a \in V_A} P_{\text{CC}}(t|a, s)P_{\text{CC}}(a|s)$. Using the entire vocabulary V_A introduces a lot of noise [7]. Instead, we use the *translingual theme*, which is a more focused and reliable indicator of topical relatedness. In addition, if we assume that T is independent of S given A, we get $P_A(t|s) \triangleq \sum_{a \in TLT_A(t,s)} P_{\text{CC}}(t|a)P_{\text{CC}}(a|s)$. [10] The independence assumption means that we are no longer constrained to use a multilingual corpus, but

[7] Here, L_S=Kannada and L_T=Hindi.

[8] We abbreviate $P_{\text{CC}}(T = t|S = s)$ to $P_{\text{CC}}(t|s)$.

[9] This is different from $P_{\text{MT}}(t|s)$, the probability that a translator would consider that t is a translation of s, which is usually used in machine translation literature [49].

[10] While this equation looks identical to the triangulation equation [43], the underlying probabilistic model there is $P_{\text{MT}}()$ (see Footnote 9), while in our case it is $P_{\text{CC}}()$.

can use several bilingual corpora—one for each language pair. This is critical, since multilingual corpora are far more difficult to obtain than bilingual corpora. Also, if the word a is not present in the L_A–L_T corpus, we need to use a non-informative uniform back-off distribution for $P(t|a)$ (as suggested by [43] for dissimilar corpora).

We use $P_A(t|s)$ as a measure of the topical similarity between t and s. In the example in Figure 1, using $P_{\{\text{en}\}}(t|s)$ along with $P_{\text{CC}}(t|s)$ results in a high value for $S_{\{\text{en}\}}(\text{अकबर}|\text{ಅಕ್ಬರ್})$, and thus improves the ranking of the translation अकबर from 6 (using S_{raw}) to 3 (using $S_{\{\text{en}\}}$).

3.3 Model for Combining Languages

Since both $P_{\text{CC}}(t|s)$ and $P_A(t|s)$ are imperfect indicators of translation correspondence, we would like to combine both scores, but weight the contribution of each distribution according to its performance on a small training set. Consequently, we chose a simple mixture model for combining information. The generative story for the model is as follows:

1. Sample a source word s uniformly from the source vocabulary V_S.
2. For each s:
   (a) Sample $j \sim Discrete(\lambda)$. (j is one of the mixture components.)
   (b) Sample $t \sim Discrete(\beta_{js})$. (A mixture component is a discrete distribution over the target vocabulary.)

Suppose we have learned, using a set of comparable corpora, the distributions $P_0(t|s) \triangleq P_{\text{CC}}(t|s)$ and $P_j(t|s) \triangleq P_{A_j}(t|s),\ j = 1 \ldots J$, for the auxiliary language set $A = \{A_j\}_{j=1}^{J}$.[11] Define

$$p(t|s,\lambda) \triangleq \sum_{j=0}^{J} \lambda_j \beta_{jst}$$

where $\beta_{jst} = P_j(t|s)$, $\lambda_j \geq 0\ \forall j$ and $\sum_j \lambda_j = 1$. Given a small training set of source-target translation pairs $\{(s_n, t_n)\}_{n=1}^{N}$[12], we can learn λ by grid search, or by maximizing the log-likelihood $\sum_n \log \sum_j \lambda_j \beta_{js_nt_n}$ w.r.t. λ.[13] For the maximum likelihood approach, we used the EM algorithm. We initialize λ randomly, and then use the following updates till convergence:

$$b_{nj} = \frac{\beta_{js_nt_n}\lambda_j}{\sum_{j'} \beta_{j's_nt_n}\lambda_{j'}}\ ,\ \lambda_j = \frac{\sum_n b_{nj}}{\sum_{j'}\sum_n b_{nj'}}\ .$$

[11] In our experiments, we have tried $J = 1, 2$ and 3.

[12] Note that this training set of a few (< 100) translation pairs is different from the seed lexicons mentioned in Section 1, which are bilingual lexicons of a few thousand translation pairs that are used by some methods (e.g. [4]) to bootstrap cross-language comparisons. We do not use such seed lexicons.

[13] We report results using the grid search in the paper, and the results using EM in the supplementary material.

We do multiple random initializations, and keep the λ with the best likelihood. Having learnt λ, we can compute $p(t|s,\lambda)$ for any word pair (s,t). The **new scoring function** $S_A()$ is defined as $S_A(t|s) \triangleq p(t|s,\lambda) = \sum_{j=1}^{J} \beta_{jst}\lambda_j$. The translation candidate t^* for s is defined as $t^* = \arg\max_t S_A(t|s)$.

Through β, other cues can also be introduced, e.g., other scoring functions on the same corpus, limited-coverage dictionaries, and multilingual WordNets.

4 Experiments and Results

We evaluated our method on 21 language pairs derived from 7 Indian languages from 2 language families—Indo-Aryan: Bengali (*bn*), Hindi (*hi*), and Marathi (*mr*), and Dravidian: Kannada (*kn*), Malayalam (*ml*), Tamil (*ta*), and Telugu (*te*). We used two auxiliary languages from different language families—Germanic: English (*en*), and Romance: French (*fr*). We extracted 35 comparable corpora (624,856 documents in total) from Wikipedia, which were the largest possible corpora possible (using all available `langlinks`). We used a state-of-the-art method for CC-TCI to measure the impact of using auxiliary languages. We also performed a user study on *WikiTSu* for the language pair Kannada-Hindi. In the remainder of this section, we refer to our method as AUX-COMB.

4.1 Experimental Setup

Corpora and Gold Standard Sets. We downloaded the Wikipedia XML dumps[14] for the 9 languages and processed them using *WiCCX*, a tool that extracts comparable corpora, cleans the documents, and restricts them to a "useful" subset of the vocabulary. The *WiCCX* tool also extracts translation pairs using `langlinks` between article titles—an approach discussed in earlier work [32]. We also create reduced gold sets for each auxiliary language set by removing words that are not present in the auxiliary corpora. Thus we obtained several gold sets $G(A)$ depending on the choice of the auxiliary language set A. The details of the corpora and gold sets are given in the supplementary material.

Evaluation Procedure. We used Monte Carlo cross-validation, which has been shown to be asymptotically consistent [50] resulting in more pessimistic predictions of performance on test data compared to normal cross-validation. The gold-standard translation pair set was divided into training and test sets in k different ways by random sampling[15]. The size of the training set (for learning λ) was fixed at d[16] for all language pairs, and the remaining translation pairs were used for testing.

[14] `http://dumps.wikimedia.org/`

[15] We fixed $k = 10$ in our experiments.

[16] We set d=40 for A={en}, {fr} and {hi}, and d=35 for A={en, fr} (proportional to the size of the gold standard set $G(A)$ available).

Given a test set in languages L_1 and L_2, for each word in L_1 in the test set, each method was used to generate a ranked list of candidate words in language L_2. Similarly, L_1 candidates were generated for L_2 words. Each ranked list was evaluated in terms of mean reciprocal rank (MRR) [51].[17] Let $tr(w)$ be the translation of w in the gold set. Given a ranked list generated for w, $RR(w) = \frac{1}{\text{Rank of } tr(w) \text{ in the list}}$. The reciprocal ranks were averaged over all words in the test set, and again averaged over all k folds in the Monte Carlo cross-validation to get the final score. Since the gold sets differed between experiments, the scores are not directly comparable. Instead, we report performance improvement over the baseline score (computed on the same gold set).[18]

Scoring Function and Baseline. Given the noisy nature of the Wikipedia corpus, we chose the **TI+Cue** method as our baseline. The TI+Cue method is a state-of-the-art method for CC-TCI, proposed in [8]. It is based on topic models [52], which work at the coarser level of topics (rather than words, or documents), and hence can be expected to smooth out noise better.[19] This method also yielded the scoring function S_{raw} (see Section 3.1) used by AUX-COMB.

For bilingual topic modeling, we used the Mallet toolbox [53] with the following configuration: regex for importing data = "`[\p{L}\p{M}]+`" (to read Unicode text with tokenization on whitespace and punctuation), Number of topics $K = \lceil \frac{\#\text{doc pairs}}{10} \rceil$, $\alpha = \frac{50}{K}$, $\beta = 0.01$ (to favor peaked distributions for topics and words [54]), Number of iterations = 1000 for estimation and 100 for inference, and Burn-in period = 100 iterations (the default settings in the toolbox).

4.2 Discussion of Results

The performance of the baseline method for $G(\{\text{en}\})$ is shown in Table 1 (left). The number in row L_S and column L_T is the performance measured when identifying translations for L_S words in language L_T. It can be seen that MRR is in the range [0.2,0.3] for most language pairs, and even lower for *bn-kn*, *kn-ml*, *kn-mr* and *ml-mr*, which have small corpora sizes (<1000). We believe that using auxiliary language corpora will be especially useful for such language pairs.

Auxiliary Languages Boost Performance. Table 1 (right) shows the improvement in MRR for AUX-COMB with English as the auxiliary language[20]. We see reasonable improvement in MRR in general, with large improvements (upto **91%**) for some language pairs. We see similar behavior with French and Hindi

[17] We also measured "Presence-at-k" (Pres@k) for k = 1 and 5. These measures showed the same trends as MRR. The details are given in the supplementary material.

[18] We report the absolute scores for the baseline on $G(\{\text{en}\})$ in Table 1 (left) to give the reader an idea of the absolute MRR scores. The absolute scores for all cases are reported in the supplementary notes.

[19] The baseline method is described in detail in the supplementary notes.

[20] We report the mean MRR across samples, and omit variances due to lack of space (e.g. the average variance was .04 for $S_{\{\text{en}\}}()$).

Table 1. *Left*: Absolute performance (in terms of MRR) of the baseline method (TI+Cue) on the English gold set $G(\{\text{en}\})$. (Poorly performing language pairs are in bold). *Right*: Percentage improvement (over baseline MRR) of AUX-COMB using $S_{\{\text{en}\}}()$. (The shading darkness of a cell is proportional to $\lambda_{\{\text{en}\}}$.)

MRR	bn	hi	kn	ml	mr	ta	te
bn	–	.3174	**.1842**	.2422	.2439	.2923	.2271
hi	.284	–	.2837	.2408	.3145	.283	.2942
kn	.2113	.2966	–	**.1273**	**.165**	.2342	.2313
ml	.2500	.3228	**.1522**	–	.2226	.2416	.2381
mr	.2230	.349	**.1403**	**.1876**	–	.2832	.2488
ta	.2731	.3232	.241	.2472	.2511	–	.2483
te	.2506	.2943	**.1748**	.3543	.2318	.2571	–

%Imp	bn	hi	kn	ml	mr	ta	te
bn	–	24.95	**90.34**	20.81	10.46	28.16	38.00
hi	7.89	–	5.71	24.09	25.02	25.97	26.14
kn	55.04	26.50	–	**91.83**	**58.55**	50.21	65.93
ml	12.32	19.08	**37.45**	–	17.74	3.93	36.67
mr	21.17	29.46	**65.93**	**39.71**	–	14.05	23.59
ta	8.46	9.41	9.67	4.81	7.81	–	21.35
te	29.49	36.94	**81.69**	19.53	33.91	42.98	–

Table 2. Percentage improvement (over baseline MRR) of AUX-COMB using $S_{\{\text{fr}\}}()$ on $G(\{\text{fr}\})$ (left), and $S_{\{\text{hi}\}}()$ on $G(\{\text{hi}\})$(right).

%Imp	bn	hi	kn	ml	mr	ta	te
bn	–	32.50	**60.04**	34.47	23.59	24.13	27.52
hi	21.37	–	22.92	31.38	8.63	18.50	19.71
kn	43.11	19.85	–	**70.15**	**32.83**	51.69	44.58
ml	22.28	16.10	**55.33**	–	11.07	28.29	43.32
mr	33.30	26.59	**49.07**	**22.73**	–	10.22	36.64
ta	33.63	11.59	24.97	21.37	7.51	–	18.22
te	20.44	18.15	**59.24**	0.74	24.32	36.07	–

%Imp	bn	hi	kn	ml	mr	ta	te
bn	–	–	**61.78**	26.08	23.37	23.79	25.32
hi	–	–	–	–	–	–	–
kn	29.79	–	–	**34.68**	**18.85**	40.08	54.47
ml	12.22	–	**72.33**	–	25.58	44.09	33.28
mr	15.15	–	**71.11**	**33.61**	–	24.24	37.81
ta	19.71	–	24.14	13.63	19.46	–	34.99
te	20.71	–	**76.78**	19.59	53.45	54.93	–

as the auxiliary language (Table 2). To show the contribution of the auxiliary language model, we shade each cell in Table 1 (right) proportional to $\lambda_{\{\mathbf{en}\}}$, the component of λ corresponding to $P_{\{\text{en}\}}$. The minimum and maximum values of $\lambda_{\{\text{en}\}}$ were 0.51 and 0.81, and the mean and median values were both 0.65.

We tried AUX-COMB with two[21] auxiliary languages to study the impact of using more languages (Table 3). The results are much better than when a single auxiliary language is used (we see upto **124%** improvement). For example, for *mr-ml*, the improvement obtained using *en* and *fr* were 39% and 22%, and using both was 83%. We see similar results for *kn-te*, *te-mr*, etc. We see robust performance for most of the 21 language pairs and for both directions.

Asymmetric Performance. As anticipated in Section 3.2, we see an asymmetry in performance for a single language pair, e.g. MRR for *te–ml* is 0.3543, while MRR for *ml–te* is 0.2381. Since the auxiliary models also have the same property, we see that the performance improvement is also not symmetric—even if the baseline performance happens to be symmetric. For example, MRR values for *ta–te* are 0.25 and 0.26, while the improvements are 21% and 42%.

[21] The model allows the inclusion of any number of auxiliary languages. However, our experimental setup requires the training pairs to be present in every auxiliary language corpus, so as to accurately measure the contribution of each auxiliary language. This restriction resulted in very small training sets when using three or more auxiliary languages, e.g. $|G(\{\text{en}, \text{fr}, \text{hi}\})| = 37$ for *kn-ml*. Due to this reason, we did not try with more auxiliary langauges for our chosen set of language pairs.

Table 3. Percentage improvement (over baseline MRR) of AUX-COMB using $S_{\{en,fr\}}()$ on $G(\{en, fr\})$.

%Imp	bn	hi	kn	ml	mr	ta	te
bn	–	36.05	**92.45**	42.59	26.95	41.55	46.90
hi	24.96	–	31.77	28.94	34.75	25.95	43.81
kn	53.36	27.27	–	**82.33**	**89.51**	52.03	94.75
ml	13.98	22.83	**51.72**	–	23.26	18.77	68.03
mr	32.10	35.66	**95.94**	**83.78**	–	12.48	42.36
ta	39.64	17.78	23.22	15.50	19.12	–	45.39
te	33.60	38.21	**124.54**	10.24	70.74	55.37	–

Table 4. Examples: for each source *kn* word, we generate the translation correspondence using TI+Cue, and using AUX-COMB (with $S_{\{en,fr\}}$) and show (a) the top-ranked *te* word, and (b) the rank of the *te* translation.

Source word		**TI+Cue**			$S_{\{\mathbf{en,fr,hi}\}}$		
kn word	Meaning	*te* word at rank 1	Meaning	Rank of transln.	*te* word at rank 1	Meaning	Rank of transln.
ಎಲೆಕ್ಟ್ರಾನ್	electron	చక్కెర	sugar	20	ఎలక్ట్రాన్	electron	1
ರಸವಿದ್ಯೆ	chemistry	గ్రీకీక్	Greek	24	శాస్త్రం	science	3
ಶನಿ	saturn	సాహిత్యము	literature	9	శని	saturn	1
ಶಿಲೀಂಧ್ರ	fungus	పర్యావరణ	environment	32	లైకెన్	lichen	4
ಗುರುತ್ವ	gravitation	గురుత్వాకర్షణ	gravitation	1	కృష్ణ	dark	3
ಜೀವಸತ್ವಗಳು	vitamins	వ్యాధి	disease	55	విటమిన్	vitamin	1
ಜೀವವೈವಿಧ್ಯ	biodiversity	పర్యావరణ	environment	9	పర్యావరణ	environment	4
ಗುಲಾಮಗಿರಿ	slavery	కౌన్సిల్	council	2	బానిసత్వం	slavery	1

Examples from kn-te. Table 4 shows some examples for *kn-te*. For each *kn* word, we take the translation correspondences using TI+Cue and AUX-COMB (with $S_{\{en,fr\}}()$) and show the *te* word at rank 1 and the rank of the correct *te* translation. We found that the top-ranked terms from both approaches were topically related but the translation was not usually at rank 1. However, AUX-COMB is able to use additional evidence from multiple languages and boost the probability of the translation so that it is ranked higher.

4.3 Further Analysis for AUX-COMB

Small Training Sets are Enough. We analyzed how sensitive our method was to the size of the training set used for learning λ. We chose the language pair *mr-te* since it had a sufficiently large gold set to allow training set size ablation, and sufficiently high performance to allow both positive and negative variation. In Figure 3 (left), we see the performance of AUX-COMB for different training set sizes. The overall trend suggests a very gradual increase in performance as training set size increases. For just 10 pairs, the performance is nearly as good as the performance for 70 pairs. The trend for *te-mr* was very similar.

Both Rare and Frequent Words Do Better. We analyzed how our method performed on words with different collection frequencies. For the language pair

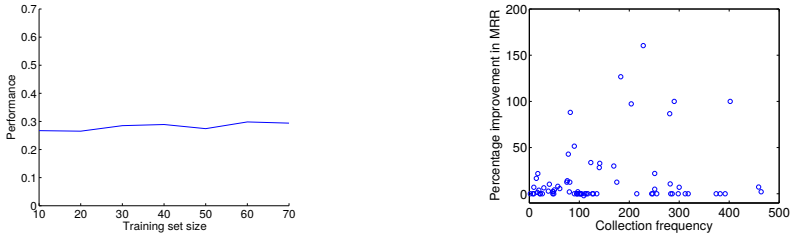

Fig. 3. *Left*: MRR for different training set sizes for *mr*-*te*. *Right*: Improvement in MRR for *te* terms with different collection frequencies, for *te*-*mr* with $S_{\text{en}}()$.

te-*mr*, we plotted the collection frequency of *te* words vs. percent improvement in MRR (Figure 3 (right)). We observe improvement over a wide range of frequencies, suggesting that the method is suitable for both rare as well as frequent words. The observations were similar for *mr* terms as well. We performed similar analyses for other term properties, viz. document frequency and average document count, and observed similar behavior.

4.4 Wikipedia Title Suggestion—User Study

We performed a user study on the *WikiTSu* system for the language pair Kannada-Hindi to assess the quality of the cross-lingual titles suggested. The quality of suggestions for source words that are Wikipedia titles has been studied in Section 4.2. In the user study, we focused on source words that are *not* Wikipedia titles. Since the Kannada Wikipedia (∼14,000 articles) is much smaller than the Hindi Wikipedia (∼100,000 articles), we chose Kannada as the source language.

Study Methodology. We randomly selected 3200 words from the *kn* corpus that were not titles, and removed common verbs, adjectives, parts of names, very common nouns, and noise words—these are unlikely to be article titles in *hi* (or any other language), giving a final list of 512 words. For each *kn* word k, we scored the *hi* vocabulary, and presented to the user the top-scoring *hi* word h that is also a Wikipedia title, with the following instructions: Suppose the user sees k in an article, and wants to know more about the concept K represented by the word k. Let H be the article corresponding to h. Score h as 1 if H is about the concept K, 0.5 if H contains information about concept K, and 0 otherwise. The above exercise was performed independently by two users.

Results. For each scoring method (TI+Cue and AUX-COMB), for each k, we averaged the relevance score given by the two users, and then averaged that over all k. The results (Table 5 (left)) show that using AUX-COMB leads to a **20% improvement** in the quality of titles. The Cohen's κ agreement between the users is good, but does not take the ordering the scores into account—a

Table 5. User study on *WikiTSu*: Average relevance score of suggested titles and user agreement metrics (left), and the weight matrix for weighted κ (right).

	TI+Cue	AUX-COMB
Avg. relevance score	0.298	**0.360**
Agreement	83%	81%
Cohen's κ	0.69	0.68
Weighted κ	0.83	0.81

		User 2		
	W	1	0.5	0
User 1	1	0	1	3
	0.5	1	0	1
	0	3	1	0

disagreement of 0 *vs.* 1 is worse than 0 *vs.* 0.5. We computed the weighted κ [55] using the weight matrix W [22] shown in Table 5 and found very good agreement.

5 Conclusions and Future Work

In this paper, we explored using auxiliary language corpora for CC-TCI. Using no resources other than comparable corpora, we demonstrated remarkable improvements in performance for 21 language pairs and applied the method to the crosslingual Wikipedia title suggestion task. This study raises interesting questions regarding the effect of the number of languages, language family, and corpus characteristics and quality. The model combination framework allows easy introduction of other cues besides auxiliary language corpora, e.g. transliteration models for names. We plan to explore these ideas in future work.

Acknowledgements. We thank Srivaths Ranganathan for the initial experiments, and Chaitra Shankar for help with the annotation. This work was supported by grants from Infosys Technologies Ltd. and the Department of Science and Technology, Government of India.

References

1. Schafer, C., Yarowsky, D.: Inducing translation lexicons via diverse similarity measures and bridge languages. COLING-02 (2002)
2. Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Djean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. ACL '04 (2004)
3. Andrade, D., Tsuchida, M., Onishi, T., Ishikawa, K.: Translation acquisition using synonym sets. In: NAACL-HLT. (2013)
4. Tamura, A., Watanabe, T., Sumita, E.: Bilingual lexicon extraction from comparable corpora using label propagation. In: EMNLP-CoNLL. (2012)
5. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. In: ACL-HLT. (2008)
6. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. ULA '02 (2002)

[22] W_{ab} is the penalty when a title is given the score a by User 1, and b by User 2.

7. Ismail, A., Manandhar, S.: Bilingual lexicon extraction from comparable corpora using in-domain terms. In: COLING. (2010)
8. Vulić, I., De Smet, W., Moens, M.: Identifying word translations from comparable corpora using latent topic models. In: ACL-HLT. (2011)
9. Udupa, R., Khapra, M.: Improving the multilingual user experience of wikipedia using cross-language name search. HLT '10 (2010)
10. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M.S., Gergle, D.: Omnipedia: bridging the wikipedia language gap. In: CHI. (2012)
11. Rapp, R.: Identifying word translations in non-parallel texts. ACL '95 (1995)
12. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. UAI (2009)
13. Lee, L., Aw, A., Zhang, M., Li, H.: Em-based hybrid model for bilingual terminology extraction from comparable corpora. COLING '10 (2010)
14. Prochasson, E., Fung, P.: Rare word translation extraction from aligned comparable documents. HLT '11 (2011)
15. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostay, T.: Term extraction, tagging, and mapping tools for under-resourced languages. In: TKE. (2012)
16. Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., Schütze, H.: A linguistically grounded graph model for bilingual lexicon extraction. In: COLING. (2010)
17. Qian, L., Wang, H., Zhou, G., Zhu, Q.: Bilingual lexicon construction from comparable corpora via dependency mapping. In: COLING. (2012)
18. Delpech, E., Daille, B., Morin, E., Lemaire, C.: Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In: COLING. (2012)
19. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: HLT-NAACL. (2009)
20. Shao, L., Ng, H.T.: Mining new word translations from comparable corpora. COLING '04 (2004)
21. Déjean, H., Gaussier, É., Sadat, F.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: COLING. (2002)
22. Rapp, R.: Automatic identification of word translations from unrelated english and german corpora. ACL '99 (1999)
23. Laroche, A., Langlais, P.: Revisiting context-based projection methods for term-translation spotting in comparable corpora. In: COLING. (2010)
24. Morin, E., Daille, B., Takeuchi, K., Kageura, K.: Brains, not brawn: The use of \smart\ comparable corpora in bilingual terminology mining. ACM Trans. Speech Lang. Process. (2008)
25. Rubino, R., Linarès, G.: A multi-view approach for term translation spotting. In: CLITP. (2011)
26. Fišer, D., Ljubešic, N.: Bilingual lexicon extraction from comparable corpora for closely related languages. In: RANLP. (2011)
27. Mausam, Soderland, S., Etzioni, O., Weld, D.S., Skinner, M., Bilmes, J.: Compiling a massive, multilingual dictionary via probabilistic inference. ACL '09 (2009)
28. Kaji, H., Tamamura, S., Erdenebat, D.: Automatic construction of a japanese-chinese dictionary via english. In: LREC. (2008)
29. Udupa, R., Saravanan, K., Kumaran, A., Jagarlamudi, J.: Mint: a method for effective and scalable mining of named entity transliterations from large comparable corpora. EACL '09 (2009)

30. Li, L., Wang, P., Huang, D., Zhao, L.: Mining english-chinese named entity pairs from comparable corpora. TALIP (2011)
31. Ji, H.: Mining name translations from comparable corpora by creating bilingual information networks. BUCC '09 (2009)
32. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: Improving the extraction of bilingual terminology from wikipedia. ACM TMCCA (2009)
33. Fung, P.: Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. VLC (1995)
34. Vulić, I., Moens, M.F.: Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In: NAACL-HLT. (2013)
35. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In: COLING. (2010)
36. Su, F., Babych, B.: Development and application of a cross-language document comparability metric. In: LREC. (2012)
37. Shezaf, D., Rappoport, A.: Bilingual lexicon generation using non-aligned signatures. ACL '10 (2010)
38. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. KDD '02 (2002)
39. Borin, L.: You'll take the high road and i'll take the low road: using a third language to improve bilingual word alignment. In: COLING. (2000)
40. Mann, G.S., Yarowsky, D.: Multipath translation lexicon induction via bridge languages. NAACL '01 (2001)
41. Tsunakawa, T., Okazaki, N., ichi Tsujii, J.: Building bilingual lexicons using lexical translation probabilities via pivot languages. In: LREC. (2008)
42. Wu, H., Wang, H.: Pivot language approach for phrase-based statistical machine translation. Machine Translation (2007)
43. Cohn, T., Lapata, M.: Machine translation by triangulation: Making effective use of multi-parallel corpora. In: ACL. (2007)
44. Utiyama, M., Isahara, H.: A comparison of pivot methods for phrase-based statistical machine translation. In: HLT-NAACL. (2007)
45. Kumar, S., Och, F.J., Macherey, W.: Improving word alignment with bridge languages. In: EMNLP-CoNLL. (2007)
46. Khapra, M.M., Kumaran, A., Bhattacharyya, P.: Everybody loves a rich cousin: an empirical study of transliteration through bridge languages. HLT '10 (2010)
47. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. ACL '05 (2005)
48. Kim, W., Khudanpur, S.: Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. ACM Transactions on Asian Language Information Processing (TALIP) **3** (2004) 94–112
49. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. (1993)
50. Picard, R.R., Cook, R.D.: Cross-validation of regression models. JASA (1984)
51. Voorhees, E.M., et al.: The trec-8 question answering track report. In: TREC. (1999)
52. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. EMNLP '09 (2009)
53. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
54. Heinrich, G.: Parameter estimation for text analysis. Technical report (2009)
55. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin (1968)