An Iterative approach to extract dictionaries from Wikipedia for under-resourced languages

Rohit Bharadwaj G
SIEL, LTRC
IIIT Hyd
bharadwaj@research.iiit.ac.in

Niket Tandon
Databases and Information Systems
Max-Planck-Institut fr Informatik
ntandon@mpi-inf.mpg.de

Vasudeva Varma
SIEL, LTRC
IIIT Hyd
vv@iiit.ac.in

Abstract

The problem of extracting bilingual dictionaries from Wikipedia is well known and well researched. Given the structural and rich multilingual content of Wikipedia, a language independent approach is necessary for extracting dictionaries for various languages more so for under-resourced languages. In our attempt to mine dictionaries for under-resourced languages, we developed an iterative approach to construct parallel corpus for building a dictionary, for which we consider several kinds of Wikipedia article information like title, infobox information, category, article text and dictionaries already built at each phase. The average precision over various datasets is encouraging with maximum precision of 76.7%, performing better than existing systems. As no language-specific resources are used, our method is applicable to any pair of language with special focus on under-resourced languages and hence breaking the language barrier.

1 Introduction

The World-Wide Web (W3) was developed to be a pool of human knowledge, which would allow collaborators in remote sites to share their ideas and all aspects of a common project ((Wardrip-Fruin and Montfort, 2003)). But at present, it is far from achieving this vision. For someone who reads only German, it is German-wide-web and Hindi reader sees it as Hindi-wide-web. To bridge the gap between the information available and languages known, Cross Language Information Access (CLIA) systems are vital. More so in these days where large content in many languages is generated daily in the form of news articles, blogs etc. Similar argument can be made on Wikipedia articles too.

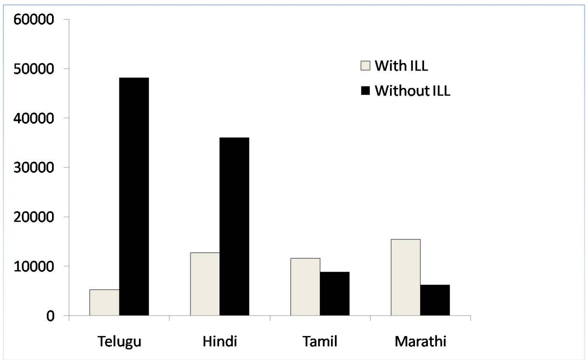


Figure 1: Number of Wikipedia pages(Y-axis) with and without Inter language link (ILL) to english in each language(X-axis)

The statistics in the figure show the growth of Wikipedia in each language irrespective of the presence of an English counterpart and we can conclude that a cross lingual system is necessary to process this rich language content. The first step in leveraging this rich multilingual resource is to build bilingual dictionaries for further processing. For resource-rich languages like French, we can afford to use various language resources like POS tagger, chunker and other prior information for extracting dictionaries but for under-resourced languages like Telugu, we need a language-independent approach for building the dictionary. Though the problem of extracting dictionaries from Wikipedia has been well researched, language-independent extraction of dictionaries is of prime importance for under-resourced languages.

Our work is different from existing techniques in the following ways:

- Existing techniques exploit Wikipedia individually in each method and combine them in the end to extract dictionaries. We, instead, use the already built dictionary iteratively to improve the precision and coverage.

Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

- Infoboxes are the vital source of information which mostly cover the proper nouns related to the title. In order to increase the coverage of proper nouns, we harnessed infoboxes of the articles.

- Not much of the work has been done in mining Wikipedia for building English-Hindi dictionary as Hindi being an under-resourced language.

As all 272 languages present in Wikipedia [1] cannot have the same resources, we aim to build bilingual dictionaries using the richly linked structure of Wikipedia documents instead of language-specific resources. The stop words are determined based on the word frequency in Wikipedia. We use English Wikipedia as base and build English-Hindi and English-Telugu dictionaries. This paper discusses building the English-Hindi bilingual dictionary. The same methodology is also applied to build English-Telugu bilingual dictionary.

2 Related Work

Dictionary building can be classified into two approaches, manual and automatic. For manual construction of dictionaries, there are various attempts like JMdict/EDCIT project (Breen, 2004). However, large amount of resources are required to construct and maintain manually built dictionaries. Also dictionaries built manually are ineffective for most of the recent vocabulary that is being added into the language everyday. In order to reduce the manual effort and keep dictionaries up-to-date, much focus is on automatic extraction of dictionaries from parallel or comparable corpora. The manually created language-specific rules, which formed the basis for automatic dictionary extraction in initial stages, were later replaced by statistical models. Initial works like (Brown et al., 1990) and (Kay and Roscheisen, 1993) were in the direction of statistical machine translation while their models can also be used for not just translation but also for dictionary building. The major requirement for using statistical methods is the availability of bilingual parallel corpora, which again is limited for under-resourced languages.

The coverage of dictionaries built using statistical translation models is also less. Though they work well for high frequency terms, they fail when terms that are not present in the corpus are encountered like technical jargaon etc. Factors like sentence structure, grammatical differences, availability of language resources and the amount of parallel corpus available further hamper the recall and coverage of the dictionaries extracted. (Fung and McKeown, 1997) showed that parallel corpus is not sufficient for good accuracy as large amount of text is added or omitted for the sake of conveying the meaning clearly.

After parallel corpora, few attempts have been made in the direction of building bilingual dictionaries using comparable corpora. (Sadat et al., 2003) uses comparable corpora along with linguistic resources for dictionary building. They perform bi-directional translation for achieving translation candidates that are later re-ranked by various methods using WWW, comparable corpora and interactive mode for phrasal translation. (Fung and McKeown, 1997) shows various ways for translating technical terms using noisy parallel corpora. Few works that use Wikipedia to construct dictionaries are described below.

(Tyers and Pienaar, 2008) built a bi-lingual dictionary using only the inter-language links(ILL). They collected a source word list and used the ILL for each word in the wordlist to find the corresponding translation. We follow similar approach for building dictionary using titles.

(Erdmann et al., 2008) extracted bi-lingual terminology from Wikipedia using ILL, Redirect pages and link text to consider the translation candidates that were ranked using the number of backward links. Though they achieved good results, they manually fixed the weights assigned to each of the category. Later, a classifier was employed to determine these weights in (Erdmann et al., 2009).

Along with ILL, we exploit infobox, Wiki categories and text of the inter language linked articles to increase the size of dictionary. The motivation behind using infobox is the fact that infobox contain the factoid information of the Wikipedia article and hence most of the query terms associated with the topic can be translated. Categories and the first paragraph of the Wikipedia article also form the prospective query terms that can be associated with the topic. Hence we had built the dictionary using them. Though we considered ILL links; Redirect text and anchor text are not used because we have considered the first few lines

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

of the article which generally contains the redirect text. We ignored anchor text as we have restricted ourselves to the first paragraph of the article instead of entire text. (Adafre and de Rijke, 2006) extracted near parallel sentences from the cross-language linked articles. They achieved the task using an available dictionary and link structure of Wikipedia. They built a cross-lingual lexicon dictionary using the ILL links present across the articles. Using this dictionary, they translated sentences in one language to other and measured the similarity using Jaccard Similarity coefficient. They compare and conclude that both the approaches give almost same similarity scores in terms of precision and recall. We are more interested in the link structure as it harnesses Wikipedia to find near-parallel sentences, that can be used to build the dictionary. We intend to use a similar method for extracting parallel sentences across the languages albeit, the dictionary we use is a more enlarged version which not only includes the ILL but also the infobox and category information. Our methods are compared to statistical machine translation at various stages because of the language independency in it, which is our main aim.

3 Proposed Method

We follow an iterative approach for achieving the task of mining dictionaries. Though we start at the same step of using ILL titles as a parallel corpus like most of the works cited in Section 2, we move into the category information, then infobox information and then into the actual text of these articles to increase the coverage of dictionaries.

Figure 2 depicts the iterative nature of our approach.

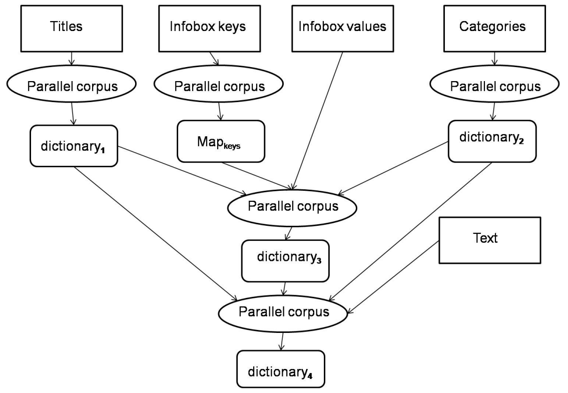
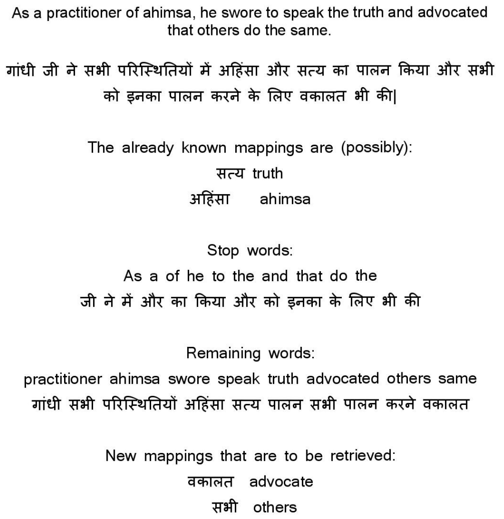


Figure 2: Iterative nature of the approach

Figure 3 shows a scenario that can occur in building dictionary from infobox values or text where we already have prior mappings between few word pairs obtained from the previous steps.



As a practitioner of ahimsa, he swore to speak the truth and advocated that others do the same.

गांधी जी ने सभी परिस्थितियों में अहिंसा और सत्य का पालन किया और सभी को इनका पालन करने के लिए वकालत भी की|

The already known mappings are (possibly):

सत्य	truth
अहिंसा	ahimsa

Stop words:

As a of he to the and that do the

जी ने में और का किया और को इनका के लिए भी की

Remaining words:

practitioner ahimsa swore speak truth advocated others same

गांधी सभी परिस्थितियों अहिंसा सत्य पालन सभी पालन करने वकालत

New mappings that are to be retrieved:

वकालत	advocate
सभी	others

Figure 3: An example of mappings

The example is an intermediate scenario of the model. Given two parallel sentences, we have the mappings of *truth* and *ahimsa* with the help of previous dictionaries. Sentences in *Remaining words* form the parallel corpus and are used to build the dictionary. The words in *New mappings* should be the dictionary formed.

In the following subsections, we explain the preprocessing steps that are followed and our approach to build dictionary from each information type of the Wikipedia article.

3.1 Preprocessing

We removed the stop words present in the lines generated for building parallel corpus using term frequency across Wikipedia. Also the text considered from English articles is converted to lower case to make the lexicons case-insensitive, since this does not affect our dictionaries.

3.2 Dictionary Building

The dictionary building process varies for each of title, category, infobox and text of the article. The common step is to generate parallel corpus, use previously built dictionaries to eliminate words and calculate the score for the pair of words. The process involved in generating the parallel text and dictionary building for each type of text (like titles, categories, infobox) is described in the following sections.

3.2.1 Titles

Each article in the English Wikipedia is considered to check if it has a cross lingual link to Hindi article. All such articles are filled in the map (where key value pairs are the titles of articles that have cross lingual link). The same is followed by considering Hindi Wikipedia and the map is updated with new title pairs. After performing the preprocessing described in Section 3.1 on the map, we form the parallel corpus for building the dictionary. The score of each word pairs is calculated by the Formula 1

$$score(w_E^i, w_H^j) = \frac{W_E^i \bigcap W_H^j}{W_E^i} \qquad (1)$$

Where w_E^i is the i^{th} word in English wordlist; w_H^j is the j^{th} word in Hindi wordlist; $W_E^i \bigcap W_H^j$ is the count of co-occurrence of w_E^i and w_H^j in the parallel corpus and; W_E^i is the count of occurrences of the word w_E^i in the corpus. After analyzing the top scored words, we found that a threshold of top 5 words contain the translation for majority of words. Hence top-5 scored Hindi words along with their scores are indexed to form $dictionary_1$.

3.2.2 Info box

The prior step of building dictionaries from Infobox involves the process of extraction of infobox. Though infobox of English articles have specific format, infobox of other under-resourced languages (e.g. Hindi, Telugu in our case) are not so well defined. We identified patterns for matching infoboxes in the articles and extracted them. Building dictionary from infobox is a two step process. The first step is to create a map of keys across the languages so that values can be mapped to create the dictionary. In the first step, the articles from the map constructed in Section 3.2.1 are considered and infobox of each article is extracted. Since Wikipedia has restriction on the way infobox keys are represented, a map of keys is constructed across the languages. Given that keys are mapped, corresponding values are mapped then. This forms the second step. We reduce the number of words in the value pair by removing the highly scored word pairs in $dictionary_1$. After this step, preprocessing described in Section 3.1 is performed and the remaining value pairs form the parallel file. The formula 1 is applied. A similar analysis like in Section 3.2.1 to find the threshold and top 10 words are indexed, which forms $dictionary_2$.

3.2.3 Categories

The process for titles is repeated here. Instead of titles pairs of cross language linked articles, we consider the categories of the articles. The only difference is the number of words that are indexed. After performing the analysis, the threshold for categories is 10. Hence top-10 words are used to build the index and the index forms the $dictionary_3$.

3.2.4 Parallel Text

The initial lines of a Wikipedia article (that generally summarizes the entire article) are considered between the articles linked by cross-lingual link. We consider lines till the first heading is encountered i.e. intuitively, we consider the abstract of the article. The parallel corpus is generated from near parallel sentences that are extracted from the introduction lines of the cross language linked articles. Similarity between each line of English text with each line of Hindi text is calculated using Jaccard Similarity coefficient by using previously built $dictionary_1$, $dictionary_2$ and $dictionary_3$. The candidate English and Hindi sentences for which the similarity coefficient is maximum are considered to be near parallel and are used to build the parallel file. We remove the already existing mappings between the pair of lines by using the dictionaries and perform the preprocessing described in Section 3.1 before populating the parallel file. The same formula 1 is used and the score for each word pair is calculated. On analysis the top 2 scored words contain the translation and hence top 2 words with scores are indexed to form $dictionary_4$.

4 Results

4.1 Dataset

Three sets of 300 English words each are generated from existing English-Hindi and English-Telugu dictionaries. These 300 words contain a mix of most frequently used words to less frequently used words. Frequency of the words is determined based on the Hindi and Telugu language news corpus. The words are POS tagged to perform the tag based analysis.

4.2 Evaluation

Precision and recall are calculated for the dictionaries built. Precision, which measures accuracy,

	Automated Eval		**Manual Eval**		**Title based Eval**	
Method	Precision	Recall	Precision	Recall	Precision	Recall
Set 1	0.464	0.554	0.777	0.554	0.570	0.434
Set 2	0.497	0.537	0.783	0.537	0.584	0.417
Set 3	0.503	0.557	0.743	0.557	0.633	0.427

Table 1: Manual and Automatic Evaluations (Hindi)

	Automated Eval		**Manual Eval**		**Title based Eval**	
Method	Precision	Recall	Precision	Recall	Precision	Recall
Set 1	0.117	0.17	0.411	0.17	0.353	0.056
Set 2	0.093	0.143	0.441	0.143	0.235	0.056
Set 3	0.117	0.2	0.441	0.143	0.285	0.07

Table 2: Manual and Automatic Evaluations (Telugu)

is the total number of words that are correctly mapped to the total number of words for which mapping exist in the test set. Precision is

$$\frac{|ExtractedCorrectMappings|}{|AllExtractedMappings|} \quad (2)$$

Recall, which measures coverage, is the fraction of the test set for which a mapping exist. Recall is

$$\frac{|ExtractedCorrectMappings|}{|CorrectMappingsUsingAvailableDictionary|} \quad (3)$$

The evaluation is both manual and automatic because one English word can map to various words in hindi. Similarly hindi word can have various spellings (due to different characters) and as we are not using language-specific resources, root word extraction and different word forms (based on sense, POS tags etc) are not be extracted. Hence manual evaluation is required to judge the correctness of the result. For manual evaluation, the results are marked either right or wrong by native speakers whereas in automatic evaluation, we check if the result word is same as returned by any other available dictionary for the language pair.

4.3 Empirical Results

The precision and recall for the 3 test sets for English-Hindi are listed in Tables 1 and 2.

As mentioned in section 4.2, manual evaluation has been carried out to check the precision. Corresponding values are also listed in Tables 1 and 2

The precision when evaluated manually is high compared to automatic evaluation. This can be attributed to few factors like

1. Various context-based translation for a single English word.
2. Different word form of the word returned and that present in the dictionary.
3. Different spelling for the same word. (different characters)

The results for the same datasets with baseline dictionary built only using the titles are listed in Tables 1 and 2.

The recall is low when using titles. The precision achieved by manual evaluation is very high compared to that of baseline dictionary. The various F-Measures for each dataset using 3 methods are listed in Tables 3 and 4.

The principle goal of using Wikipedia is to cover large terminology and hence recall is as important as precision. In such cases, our method outperforms the baseline dictionary.

In case of proper nouns, where existing MT methods fail due to unavailability of parallel corpus, our method gives encouraging results, as listed in Table 5.

Precision	Recall	F-Measure(F_1)
0.715	0.787	0.749

Table 5: Accuracy for proper nouns (Hindi)

The coverage of dictionaries can be estimated by considering the number of words that are present in the dictionary. Figure 4 details number of unique words(Y-axis) that are freshly added from each category(X-axis). The maximum word count for bilingual pairs in (Tyers and Pienaar, 2008) is 4913, which is very less when compared with our system. Though the number of articles and the cross-language links affect the coverage of the dictionaries, our system performs considerably well in extracting biligual word pairs (dictio-

	Set-1			Set-2			Set-3		
Method	$F_{0.5}$	F_1	F_2	$F_{0.5}$	F_1	F_2	$F_{0.5}$	F_1	F_2
Baseline dictionary	0.536	0.492	0.455	0.541	0.486	0.442	0.577	0.509	0.456
Automatic Evaluation	0.474	0.505	0.533	0.504	0.516	0.528	0.513	0.528	0.545
Manual Evaluation	0.719	0.646	0.587	0.717	0.637	0.573	0.696	0.636	0.586

Table 3: F-Measure Accuracy (Hindi)

	Set-1			Set-2			Set-3		
Method	$F_{0.5}$	F_1	F_2	$F_{0.5}$	F_1	F_2	$F_{0.5}$	F_1	F_2
Baseline dictionary	0.172	0.097	0.068	0.144	0.091	0.066	0.176	0.112	0.082
Automatic Evaluation	0.125	0.139	0.156	0.100	0.112	0.129	0.127	0.147	0.175
Manual Evaluation	0.320	0.249	0.192	0.311	0.216	0.165	0.311	0.216	0.165

Table 4: F-Measure Accuracy (Telugu)

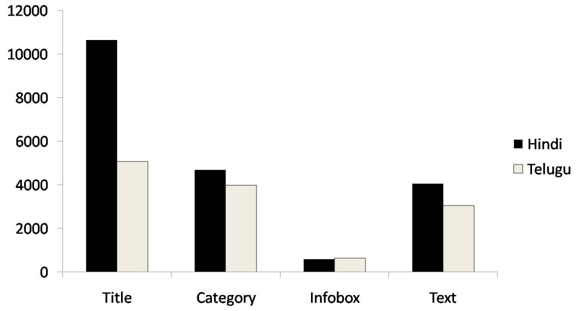


Figure 4: Coverage of Dictionaries(Y-axis) with respect to each information type(X-axis)

nary). It is found that dictionaries generated by titles and infobox are more accurate than the dictionaries constructed using categories and text. Another interesting observation is that the word pairs generated from categories and text, though, not accurate, are related to the query word. This feature can be exploited in CLIA systems where query expansion in the target language plays a vital role.

The final summary of the results attained are listed in Tables 3 and 4.

4.4 Comparison with other existing dictionary building systems

Table 6 compares our method with (Erdmann et al., 2008), who built a English-Japanese dictionary using Wikipedia using various approaches. The results considered from their work are the precision and recall their system attained for high and low frequency words. For high frequency words, they tuned their system to achieve high precision whereas for low frequency words, their system concentrated on achieving high recall.

Our results are the average precision and recall over the 3 datasets. These statistics shows that our method performs on par with the existing methods of automatic multilingual dictionaries using Wikipedia.

Approach	**Precision**	**Recall**
Existing(High precision)	0.781	0.225
Existing(High recall)	0.333	0.613
Our approach(Hindi)	0.767	0.549

Table 6: Comparing the results with other existing approaches

5 Conclusion and future work

We have described the automatic construction of a bilingual dictionary over Wikipedia, by an iterative method. Overall, the construction process consists of identifying near comparable corpora from title, infobox, category and text of the articles linked by ILL. Our focus has been low resourced languages. The system is extensible considering it is language-independent and no external language resource has been used.

We are working to apply the system to query building in CLIA systems. Before moving beyond Wikipedia, we want to consider entire text of the Wiki-article and other structure information like headings and anchor text to enhance the coverage of the dictionary. Further, multi-word translation is also envisioned. Our work will be of particular importance to the under-resourced languages where the comparable corpora are not easily obtainable.

References

S.F. Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. *NEW TEXT Wikis and blogs and other dynamic text sources*, page 62.

J. W. Breen. 2004. JMdict:A Japanese-Multilingual Dictionary. In *COLING Multilingual Linguistic Resources Workshop*, pages 71–78.

P.F. Brown, J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):85.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. In *Database Systems for Advanced Applications*, pages 380–392. Springer.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2009. Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(4):1–17.

P. Fung and K. McKeown. 1997. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1):53–87.

M. Kay and M. Roscheisen. 1993. Text-translation alignment. *computational Linguistics*, 19(1):121–142.

F. Sadat, M. Yoshikawa, and S. Uemura. 2003. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 141–144. Association for Computational Linguistics.

F.M. Tyers and J.A. Pienaar. 2008. Extracting bilingual word pairs from Wikipedia. *Collaboration: interoperability between people in the creation of language resources for less-resourced languages*, page 19.

N. Wardrip-Fruin and N. Montfort. 2003. *The NewMediaReader*. The MIT Press.