Language-Independent Context Aware Query Translation using Wikipedia

Rohit Bharadwaj G
Search and Information Extraction Lab
LTRC
IIIT Hyderabad, India
`bharadwaj@research.iiit.ac.in`

Vasudeva Varma
Search and Information Extraction Lab
LTRC
IIIT Hyderabad, India
`vv@iiit.ac.in`

Abstract

Cross lingual information access (CLIA) systems are required to access the large amounts of multilingual content generated on the world wide web in the form of blogs, news articles and documents. In this paper, we discuss our approach to query formation for CLIA systems where language resources are replaced by Wikipedia. We claim that Wikipedia, with its rich multilingual content and structure, forms an ideal platform to build a CLIA system. Our approach is particularly useful for under-resourced languages, as all the languages don't have the resources(tools) with sufficient accuracies. We propose a context aware language-independent query formation method which, with the help of bilingual dictionaries, forms queries in the target language. Results are encouraging with a precision of 69.75% and thus endorse our claim on using Wikipedia for building CLIA systems.

1 INTRODUCTION

Cross lingual information access (CLIA) systems enable users to access the rich multilingual content that is created on the web daily. Such systems are vital to bridge the gap between information available and languages known to the user. Considerable amount of research has been done on building such systems but most of them rely heavily on the language resources and tools developed. With a constant increase in the number of languages around the world with their content on the web, CLIA systems

are in need. Language independent approach is particularly useful for languages that fall into the category of under-resourced (African, few Asian languages), that doesn't have sufficient resources. In our approach towards language-independent CLIA system, we have developed context aware query translation using Wikipedia. Due to voluntary contribution of millions of users, Wikipedia gathers very significant amount of updated knowledge and provides a structured way to access it.
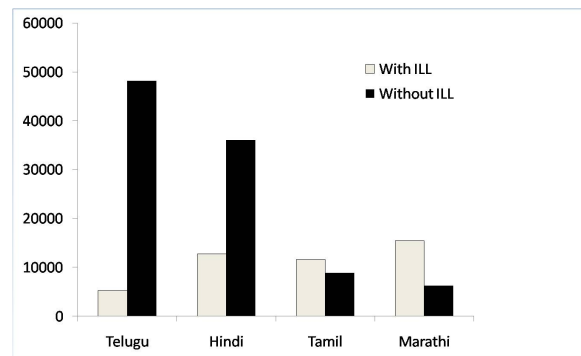

Figure 1: Number of Wikipedia pages(Y-axis) with and without Inter language link (ILL) to English in each language (X-axis)

The statistics in the Figure 1 show that it has rich multilingual content and is growing independent of the presence of English counter part. With its structurally rich content, it provides an ideal platform to perform cross lingual research. We harness Wikipedia and its structure to replace the language specific resources required for CLIA.

Our work is different from existing approaches in

terms of

- No language resource has been used at any stage of query translation.

- Wikipedia structure has been fully utilized for achieving CLIA between English and Hindi, unlike the existing approaches, especially for query formation.

We have constructed a bilingual dictionary using cross lingual links present across the articles of same topic in different languages. As each word in the dictionary can have several translations based on various attributes like context, sense etc, we need a mechanism to identify the target word accurately based on the context of the query. To identify the context of a query, "Content Words", that are built for each Wikipedia article, are used. "Content Words" of the article are similar to the tags of the article, that reflects the context of the article in a more detailed way.

In this paper, we detail our approach in forming this "Content Words" and using them to form the query. Since our approach is language-independent and context-aware, we used a metric proposed by (Bharadwaj and Varma, 2011) to evaluate along with a dicationary-based metric. The system is built between languages English and Hindi. Hindi is selected as target language because of the availability of resources for evaluation. As our approach is language-independent, it can be used to translate queries between any pair of languages present in Wikipedia. The remainder of paper is organized as follows. Section 2 shows the related work. Proposed method is discussed in Section 3. Results and Discussion are in Section 4. We finally conclude in Section 5.

2   RELATED WORK

We discuss the related work of the two stages are involved in our system of language-independent context aware query translation,

- Resource building/ collection (Dictionaries in our case)

- Query formation

Dictionary building can be broadly classified into two approaches, manual and automatic. At initial stages, various projects like (Breen, 2004) try to build dictionaries manually, taking lot of time and effort. Though manual approaches perform well, they lag behind when recent vocabulary is considered. To reduce the effort involved, automatic extraction of dictionaries has been envisioned. The approach followed by (Kay and Roscheisen, 1999) and (Brown et al., 1990) were towards statistical machine translation, that can also be applied to dictionary building. The major requirement for using statistical methods is the availability of bilingual parallel corpora, that again is limited for under-resourced languages. 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.

After parallel corpora, attempts have been made to construct bilingual dictionaries using various types of corpora like comparable corpus (Sadat et al., 2003) and noisy parallel corpus (Fung and McKeown, 1997). Though there exist various approaches, most of them make use of the language resources. Wikipedia has also been used to mine dictionaries. (Tyers and Pienaar, 2008), (Erdmann et al., 2008), (Erdmann et al., 2009) have built bilingual dictionaries using Wikipedia and language resources. We have mined our dictionaries similarly considering the cross lingual links present. Our approach to dictionary building is detailed in section 3.

Wikipedia has been used for CLIA at various stages including query formation. Most recently, Wikipedia structure has been exploited in (Gaillard et al., 2010) for query translation and disambiguation. In (Schönhofen et al., 2008), Wikipedia has been exploited at all the stages of building a CLIA system. We tread the same path of (Schönhofen et al., 2008) in harnessing Wikipedia for dictionary building and query formation. Similar to them we extract concept words for each Wikipedia article and use them to disambiguate and form the query.

For evaluation purposes, we adapted evaluation measures based on Wikipedia and existing dictionaries (Bharadwaj and Varma, 2011). The authors have proposed a classification based technique, using Wikipedia article and the inter-language links

present between them to classify the sentences as parallel or non-parallel based on the context of the sentences rather than at the syntactic level. We adopt a similar classification based technique and build feature vectors for classification using Support Vector Machines (SVM [1]) for evaluation.

3 PROPOSED METHOD

The architecture of the system is given in the Figure 2.

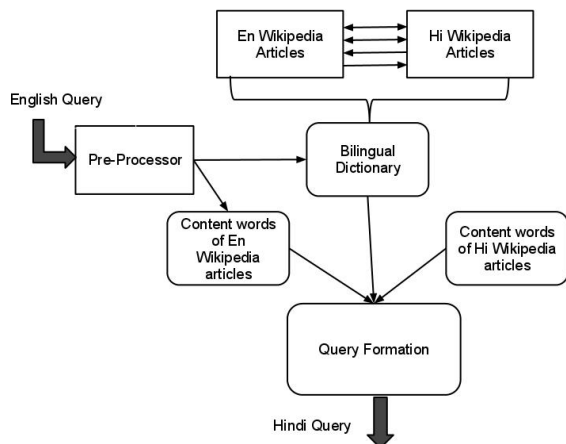

Figure 2: Architecture of the system

The following subsections describe each module in detail.

3.1 Dictionary Building

Bilingual dictionaries (English-Hindi) are built from Wikipedia by mining parallel/ near-parallel text from each structural information like title, infobox, category and abstract (initial paragraph) of the English(En) and Hindi(Hi) articles that are connected with Inter language link (ILL, arrows between En Wikipedia articles and Hi Wikipedia articles in Figure 2). The motivation for considering the other structural information of the Wikipedia article is to increase vocabulary of the dictionary both in terms of the number of words and categories of words. Titles, Infobox and Categories of the article consider only named entities that are used in the language.

[1] http://www.cs.cornell.edu/People/tj/svm_light/

To increase the coverage of the dictionary and also to include other categories of words (like negations, quantifiers etc), abstract of the article is considered. Also the Inter language links between the articles are assumed to be bi-directional even if they are uni-directional. An approach similar to (Tyers and Pienaar, 2008) is followed to construct dictionaries. The dictionary is constructed iteratively by using the previously constructed dictionaries from each structure. The structural aspects of the article used are

- Title: Titles of the articles linked.

- Infobox: Infobox of the articles that are linked.

- Category: Categories of the articles linked.

- Abstract: The inital paragraph of the articles linked are considered as the article abstracts and are used for dictionary building.

A dictionary consists of word and its several possible translations, scored according to their alignment scores. Each structural information is used to enhance the dictionary built previously. Dictionary built from titles are used as starting point. As each English word is mapped to several Hindi words, filtering of words or re-ranking of the words at query formation is vital. The scoring function used for the words while building the dictionary is

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

Where w_E^i is the ith word in English word list; w_H^j is the jth word in Hindi word list; $W_E^i \bigcap W_H^j$ 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.

3.2 Building Content words

The context of each English Wikipedia article A_i is extracted from the following structural information of the article.

- Title : Title of the article

- Redirect title : Redirect title of the article, if present.

147

- Category : Categories of the article that are pre-defined.

- Subsections : Titles of the different subsections of the article.

- In-links : Meta data present in the links to this article from other articles in same language.

- Out-links : Meta data of the links that link the current article to other articles in same language.

As these structural attributes are spread across the article, they help to identify the context (orientation) of the article in depth when compared with the Categories of the article. Each structural aspect described above have unique content that will help to identify the context of the article. "Content Words" are formed from each of these structural aspects. Word count of the words present in each of the above mentioned attributes are calculated and are filtered by a threshold to form the context words of the article. The threshold for filtering has been calculated by manual tagging with the help of language annotators. "Content Words" for the Hindi articles are also formed similarly. The formation of "Content Words" is similar to tagging but is not a stricly tagging mechanism as we have no constraint on the number of tags. Category alone can help to get the context but considering in-links, out-links, subsections will increase the depth of context words and will reduce the information lost by tagging the words.

3.3 Query formation

Query formation of our system depends on the context words built. For an English query (q_E) that contains the words w_E^i (i: 0 to n),

- Build W_H of size m, that contains the words returned by the dictionary for each of the words.

- For all words in (q_E), extract all the articles a_i^k (k: 0 to n) with w_E^i as one of its context word.

- Form the corresponding Hindi set of articles A_h

using the cross lingual link, if present in the English article set constructed in the above step.

- For each Hindi word w_H^j (j: 0 to m), add it to Hindi query (q_H) if at least one of the articles a_i (with w_H^j as its context word) is present in A_h.

This approach helps to identify the context of the query as each query is represented by a set of articles instead of query words, that forms the concepts that the query can be interpreted to limited to Wikipedia domain. Queries are translated based on the architecture described in Figure 2.

4 Results and Discussion

4.1 Evaluation, Dataset and Results

A classification based approach and a dictionary based approach are employed to calculate the accuracy of the queries translated. 400 sentences with their corresponding translations (English-Hindi) have been used as test set to evaluate the performance of the query formation. The sentence pairs are provided by FIRE[2]. These sentences contain all types of words (Named entities, Verbs etc) and will be referred to as samples. The English language sentences are used as queries and are translated to Hindi using the approach described. Before forming the query, stop words are removed from the English sentence. The query lengths after removing stop words vary from 2 words to 8 words. The dictionary used for evaluation is an existing one, Shabdanjali[3]. In the following sections, we describe our two evaluation strategies and the performance of our system using them.

4.1.1 Dictionary based evaluation

Shabdanjali dictionary has been used to evaluate the translated queries. The evaluation metric is word overlap, though it is relaxed further. The formula

---

[2]http://www.isical.ac.in/ clia/

[3]Shabdanjali is an open source bilingual dictionary that is most used between English and Hindi. It is available at `http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html`

used for calculating the precision is

$$precision = \frac{No.ofCorrectSamples}{TotalNumberofSamples} \quad (2)$$

A sample is said to be correct if its $overLapScore$ is greater than threshold instead of complete overlap. The $overLapScore$ of each sample is measured using Formula 3. Threshold is the average $overLapScore$ of the positive training set used for training the classifier (Training dataset is discussed in Section 4.1.2).

$$overLapScore = \frac{No.ofWordOverlap}{TotalNumberofWords} \quad (3)$$

The number of word overlaps are measured both manually and automatically to avoid inconsistent results due to varios syntactic representation of the same word in Wikipedia.

The precision for the test dataset using this approach is 42.8%.

4.1.2 Classification based evaluation

As described in Section 2, we have used a classification based technique for identifying whether the translated queries contain the same information or not. We have collected 1600 pairs of sentences where 800 sentences are parallel to each other (positive samples, exact translations) while the other half have word overlaps, but not parallel, (not exact translations but have similar content) form the negative samples. Various statistics are extracted from Wikipedia for each sentence pair to construct feature vector as described in (Bharadwaj and Varma, 2011). Each English and Hindi sentences are queried as bag-of-words query to corresponding Wikipedia articles and statistics are extracted based on the articles retrieved. The classifier used is SVM and is trained on the feature vectors generated for 1600 samples. The precision in this approach is the accuracy of the classifier. The formula used for calculating the accuracy is

$$accuracy = \frac{No.ofSamplesCorrectlyClassified}{TotalNumberofSamples}$$
$$(4)$$

The correctness of the sample is the prediction of the classifier. The precision for the test set is 69.75%.

4.2 Discussion

The precision achieved by classification based evaluation is higher than that of existing dictionary (Shabdanjali) primarily due to

- Dictionary (Shabdanjali) doesn't contain words of the query. (Coverage is less).

- Word forms present in the dictionary are different to that of words present in translated query. (Ex: spelling, tense etc).

To negate the effect of above factors, classification based evaluation (4.1.2) has been considered. Classification based evaluation shows that the results are better when the entire sentence and its context is considered. As there are no existing systems that translate queries based on the context and language independent, our results are encouraging to work in this direction. Since no language resources were used, our approach is scalable and can be applied to any pair of languages present in Wikipedia. The relatively low coverage of the dictionaries built using Wikipedia structure also affects the process of query translation. In future, the coverage of dictionaries can also be increased by considering other structural properties of Wikipedia.

5 Conclusion

In this paper, we have described our approach towards building a language-independent context aware query translation, replacing the language resources with the rich multilingual content provider, Wikipedia. Its structural aspects have been exploited to build the dictionary and its articles are used to form queries and also to evaluate them. Further exploitation of Wikipedia and its structure to increase the coverage of the dictionaries built will increase the overall precision. Though queries are translated in a language-independent way, using language resources of English, as it is a richly resourced language, for query formation is also envisioned.

References

Rohit G. Bharadwaj and Vasudeva Varma. 2011. Language independent identification of parallel sentences

149

using wikipedia. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 11–12, New York, NY, USA. ACM.

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.

B. Gaillard, M. Boualem, and O. Collin. 2010. Query Translation using Wikipedia-based resources for analysis and disambiguation.

M. Kay and M. Roscheisen. 1999. Text-translation Alignment. In *Computational Linguistics*, volume 19, pages 604–632.

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.

P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. 2008. Cross-language retrieval with wikipedia. *Advances in Multilingual and Multimodal Information Retrieval*, pages 72–79.

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.