Language Independent Identification of Parallel Sentences 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

This paper details a novel classification based approach to identify parallel sentences between two languages in a language independent way. We substitute the required language specific resources by the richly structured multilingual content of Wikipedia. Our approach is particularly useful to extract parallel sentences for under-resourced languages like most Indian and African languages, where resources are not readily available with necessary accuracies. We extract various statistics based on the cross lingual links present in Wikipedia and use them to generate feature vectors for each sentence pair. Binary classification of each pair of sentences into parallel or non-parallel has been done using these feature vectors. We achieved a precision upto 78% which is encouraging when compared to other state-of-art approaches. These results support our hypothesis of using Wikipedia to evaluate the parallel coefficient between sentences that can be used to build bilingual dictionaries.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing, Dictionaries, Linguistic processing

General Terms

Measurement, Languages, Algorithms

Keywords

Parallel sentences, Language Independent, Wikipedia

1. INTRODUCTION

Identification of parallel sentences is one of the major aspects in building dictionaries that affect the growth of cross lingual information access systems. Established techniques like statistical machine translation use parallel sentences to build dictionaries that help in translating the query. There are various techniques employed to build the parallel sentences from comparable text but most of the methods use language specific tools like Named entity recognizer, parser and chunker. Methods employed to compute sentence similarity and to identify parallel sentences are similar because of the common ground both the tasks share.

Word alignment techniques are used as a starting step to identify parallel sentences in [3] and [5]. Many similar approaches use either bilingual dictionary or other translation resources, for computing sentence similarity. Unavailability of language resources limits

most of the existing approaches for calculating sentence similarity. We develop a method that can substitute language resources with Wikipedia[1] and identify parallel sentences in a language independent way. Wikipedia's structural representation of the topic is helpful for various information retrieval and extraction tasks. It is used in [4] to extract parallel sentences for Statistical Machine Translation (SMT). The authors used word alignments and other lexical features along with Wikipedia to build feature vectors that are used for classification. [1] and [2] discuss different models to add translations of the articles with the help of resources mined from Wikipedia and by social collaboration respectively. Our method is particularly useful in identifying or evaluating the translations that are either human generated or machine generated like [1] and [2].

Link structure and meta data of the Wikipedia article is used to identify parallel sentences. A classification based approach is employed by building feature vectors for each pair of sentences. These feature vectors are based on the existence of cross lingual link in an article and retrieval of an article when queried using the sentence. As no language specific resources are used, our approach is scalable and can be used between any pair of 272 languages that have cross lingual links in Wikipedia. Our work is different from existing works in the following ways

1. No language specific resources are used to identify sentence similarity.
2. Wikipedia structure is exploited towards identification of parallel sentences between languages especially between English and Hindi.

2. PROPOSED METHOD

Our approach is based on the existence of cross lingual links between articles in different languages on the same topic. Sentence similarity is computed using their semantics rather than syntax. Relevant Wikipedia articles are used to obtain the information contained in the sentence. English sentence is formed into a bag of words query and queried on the English index; Hindi sentence is similarly queried. As Wikipedia contains lots of structural information, we have constructed three different types of indices to identify the importance of each structural information and also considering the time and space constraints. The entire text of English and Hindi Wikipedia articles is used to construct indices E_1 and H_1, meta data of the articles (infobox, in-links, out-links and categories) is used to construct indices E_2 and H_2 while titles and redirect titles are used to build indices E_3 and H_3. A difference

---

[1]Wikipedia (`http://www.wikipedia.org`) is a well known free content, multilingual encyclopedia written collaboratively by contributors around the world.

the government also has an important role to play

सरकार को भी एक महत्वपूर्ण भूमिका निभानी है

English Index

Hindi Index

government of India
government
bribery corruption
public school government funded

English Articles Retrieved

सरकार
भारत सरकार
भूमिका प्रकाशन
सत्यजित राय को मिले सम्मान

Hindi Articles Retrieved

list of awards conferred on satyajit ray

सार्वजनिक विद्यालय

Articles that are not retrieved

Statistics used to build feature vector are :

1) Number of English articles for which corresponding Hindi articles are retrieved and vice versa
2) Number of English articles for which corresponding Hindi articles are not retrieved and vice versa.
3) Total Number of English and Hindi articles retrieved.
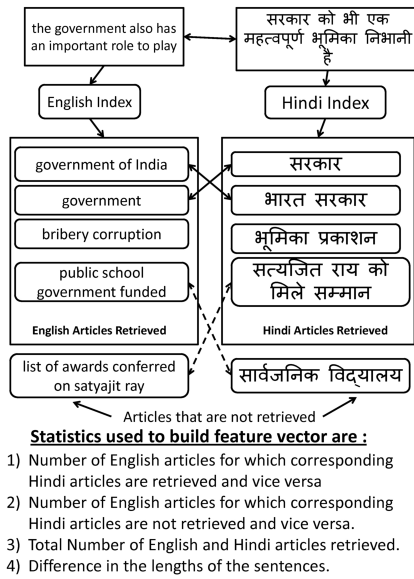4) Difference in the lengths of the sentences.

Figure 1: Example of the possible scenarios

in information represented by English and Hindi sentence is visible only when corresponding pages are not retrieved. Multilingual Wikipedia pages on the same topic need not be complete or comparable but approaches like [2] and [1] work towards creation of complete articles. As the search is with bag of words, it is assumed that all related articles will contain at least one word of the sentence. Feature vectors are built using various statistics generated (as shown in Figure:2). The classifier used is Support Vector Machines (SVM [2]) with default parameters.

3. DATASET

For training and testing, positive and negative samples are needed. Positive samples for our experiments are taken from word aligned parallel corpus provided by Language Technologies Research Center [3]. To make our classifier more generic and robust, two types of negative samples are constructed. These are constructed from the queries of FIRE dataset[4]. The sentences considered for a pair are of minimum length 4 and also of nearly same length (with a maximum difference of 2). The two types of negative samples are

$CloseNegatives$: These sentences talk about the same entity or topic and also contain word overlaps (found using a dictionary[5]), they are not parallel.

$FarNegatives$: The sentence pair talks about different entity or topic. There is no relation between the sentences either at the word level or at the information level.

The number of training and testing samples before and after removal of stop words are 1600 and 800, 800 and 400 respectively with equal distribution of positive and negative samples. The reduction in size is due to the constraints placed on sentences lengths. A baseline system is built using shabdanjali dictionary with sentence similarity computed using Jaccard index.

---

[2] http://www.cs.cornell.edu/People/tj/svm_light/
[3] http://ltrc.iiit.ac.in
[4] http://www.isical.ac.in/~fire/data_download.html
[5] 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

Indices	E_1,H_1	E_2,H_2	E_3,H_3	Baseline
Stop words removed (Close)	72.50	57.00	66.50	49.50
Stop words removed (Far)	78.00	57.50	76.50	50.50
Stop words removed (Far + Close)	75.50	57.00	74.00	50.00
Stop words not removed (Close)	66.25	49.25	56.25	49.75
Stop words not removed (Far)	74.75	56.00	59.25	51.00
Stop words not removed (Far + Close)	64.5	51.00	60.00	51.00

Table 1: Precision over different datasets and indices

4. RESULTS AND DISCUSSION

For the default parameters of SVM, the precision of our approach for different indices is calculated. The values are detailed in the Table:3. Close and Far in Table 3 are the type of negative samples used in training and testing. The results achieved are better than the baseline system considered. The under-performance of the baseline system can be attributed to various factors like the coverage of the dictionary considered, different word forms handled by the dictionary etc. Although baseline system is relaxed by considering non-zero jaccard similarity coefficient, our approach outperforms it. The best performance is obtained when the entire article is considered for indexing. The low performance for indices E_2 and H_2 is due to errors in the extraction of meta data for Hindi language.

5. CONCLUSION AND FUTURE WORK

As our approach is language independent, it can be applied to any pair of languages given that the languages exist in Wikipedia. Also the index can be from any multi lingual content that have cross lingual links like Wikipedia. We are focusing on improving the classifier quality with various other features that can be extracted by using the rich multi lingual content of Wikipedia. Later we want to construct bilingual dictionaries with the parallel sentences identified by this approach. Along with building bilingual dictionaries, it will be interesting to use this approach in various applications like cross lingual plagiarism detection and evaluation of query translations.

6. REFERENCES

[1] L. Huberdeau, S. Paquet, and A. Désilets. The Cross-Lingual Wiki Engine: enabling collaboration across language barriers. In *Proceedings of the 4th International Symposium on Wikis*, pages 1–14. ACM, 2008.

[2] A. Kumaran, K. Saravanan, N. Datha, B. Ashok, and V. Dendi. Wikibabel: A wiki-style platform for creation of parallel data. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 29–32. ACL, 2009.

[3] D. Munteanu, A. Fraser, and D. Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of HLT/NAACL*, 2004.

[4] J. Smith, C. Quirk, and K. Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of HLT/NAACL*, pages 403–411. ACL, 2010.

[5] C. Tillmann. A Beam-Search extraction algorithm for comparable data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 225–228. ACL, 2009.