Enrichment of Bilingual Dictionary through News Stream Data

Ajay Dubey*, **Parth Gupta**†, **Vasudeva Varma***, **Paolo Rosso**†

*Search and Information Extraction Lab
International Institute of Information Technology Hyderabad, India
ajay.dubey@research.iiit.ac.in, vv@iiit.ac.in

†Natural Language Engineering Lab
PRHLT Research Center
Universitat Politècnica de València, Spain
http://www.dsic.upv.es/grupos/nle
{pgupta,prosso}@dsic.upv.es

Abstract

Bilingual dictionaries are the key component of the cross-lingual similarity estimation methods. Usually such dictionary generation is accomplished by manual or automatic means. Automatic generation approaches include to exploit parallel or comparable data to derive dictionary entries. Such approaches require large amount of bilingual data in order to produce good quality dictionary. Many time the language pair does not have large bilingual comparable corpora and in such cases the best automatic dictionary is upper bounded by the quality and coverage of such corpora. In this work we propose a method which exploits continuous quasi-comparable corpora to derive term level associations for enrichment of such limited dictionary. Though we propose our experiments for English and Hindi, our approach can be easily extendable to other languages. We evaluated dictionary by manually computing the precision. In experiments we show our approach is able to derive interesting term level associations across languages.

Keywords: cross-lingual, statistical dictionary generation, news stream

1. Introduction

Cross-lingual term associations are very important for many inter-lingual applications. Bilingual dictionary is one such important resource where entries are word translations. Statistically generating a bilingual dictionary is more difficult for an under-resourced language pair for which not large amount of parallel and/or comparable data is available. Although there are few parallel corpora on the web, there is much comparable data and even more quasi-comparable. Documents in quasi-comparable collections are not topically aligned. Although our approach is generic, in this work we focus on the task of bilingual dictionary generation for an under-resourced language pair English-Hindi (en-hi).

One of the largest cross-lingual resource for en-hi pair is Wikipedia. Aligned Wikipedia for en-hi contains 30k documents. Although it is a large and highly structured resource, it is growing at a very slow rate for under-resourced languages like Hindi compared to other languages like English, Spanish, Chinese, Arabic etc.

Hindi is very under-represented on the web because of many technical and socio-cultural reasons (Ahmed et al., 2011). The only few sources of Hindi text on the web are news articles, personal blogs, social media and song lyrics. News stream data look very attractive because of its natural language content and continuous generation in large amount across languages.

In this study we try to identify the news stories across languages which are reported on the same news event and then extract cross-lingual word translations from such articles. Recently Barker and Gaizauskas (2012) have defined the taxonomy of news articles. According to the same taxonomy, we try to link news stories across languages with same focal event which might serve as a comparable corpora. Afterwards, we apply dictionary generation algorithm to extract bilingual dictionary. Our method is completely automatic and does not rely on any labelled information. We compare the dictionary generated by our method to the dictionary generated from Wikipedia. Naturally the dictionary generated from the Wikipedia is of high quality because of the highly structured information like title, section, subsection etc. and inter-lingual links. We demonstrate that our method is able to find some associations which are not present in the Wikipedia. We also present a thorough analysis of the generated dictionary.

In Section 2. we discuss the related work. We present our method in detail in Section 3.. Section 3 contains the results and analysis of our method. Finally in Section 5. we outline the main findings of the study.

2. Related Work

There have been approaches to extract bilingual content out of quasi-comparable corpus. Most of the attempts are towards extracting parallel sentences or fragments out of non-parallel or quasi-comparable corpora (Fung and Cheung, 2004; Munteanu and Marcu, 2005; Munteanu and Marcu, 2006; Chu et al., 2013). All of these approaches rely on a strong seed translation lexicon and take some type of bootstrapping approach.

The work to obtain cross-language term associations for cross-language information retrieval (CLIR) task is highly related to this approach. The early work of Sheridan and Ballerini (1996) used SDA corpus to align German-Italian news stories based on date and assigned descriptors and later cross-lingual term associations were used for the task of CLIR. Braschler and Schäuble (1998) also devised similar approach where they linked English news stories to German stories using meta descriptors and common proper nouns[1] They replaced the monolingual ranklist by the corresponding cross-lingual aligned article and then expanded the query through pseudo-relevance feedback techniques. These approaches are tested on specific task of CLIR where finding the word translations is not a constraint. In a similar approach to find word translations, Rapp (1999) carried experiments over English-German news stories with a small base German to English base lexicon and tested the accuracy over 100 test words.

The dimensionality reduction techniques are also used to align cross-lingual documents. The most celebrated approach is cross-language latent semantic indexing (CL-LSI) (Dumais et al., 1997). CL-LSI falls into the category of linear dimensionality reduction techniques which by using parallel documents tries to reduce the dimensionality to top k principal components to represent the data in reduced abstract space. The translingual documents are compared to each other in the low-dimensional abstract space and the closest pairs are considered topically aligned. There has been an extension to CL-LSI called oriented principal component analysis (OPCA) which formulates the problem as generalised eigenproblem (Platt et al., 2010). These approaches require parallel training collection for training.

Recently the cross-language !ndian news story search (CL!NSS) track of FIRE has created a benchmark collection and evaluation framework for news stories linking task for resource poor language pairs like English-Hindi (en-hi) and English-Gujarati (en-gu) (Gupta et al., 2012). The approaches based on machine translation techniques have shown to perform well but they are considerably slow especially when dealing with document collection of a few hundred thousand (Palkovskii, 2012; Gupta et al., 2012; Gupta et al., 2013, and references therein).

3. Approach

For an under-resourced language pair like en-hi, the existing attempts are not reliable as any strong parallel seed corpus is unavailable. Hence, it is important to devise a technique which can derive cross-language term translations from quasi-comparable corpora. Such approaches should have two inherent characteristics: *i)* they should be fast enough to digest large amount of data quickly (IR like approach), and *ii)* they should not depend on high-level cross-lingual resources.

The final aim of our approach is to generate a bilingual dictionary automatically from a cross-lingual quasi-comparable corpus. There are two natural component of our approach

- News stories linking
- Inflating the dictionary entries

The former step links the target article $t_i \epsilon T_{l1}$ to the source article $s_j \epsilon S_{l2}$, where T and S are the collection of target and source news stories respectively in language l_1 and l_2. The latter step takes the most frequent pairs of co-occurring words from the source and target news stories considered as comparable documents.

3.1. News Story Linking

In order to compare the documents in the collection across languages we first index each article in the source collection. The source articles are composed of title and content field. We index both fields separately and the unit of index is term and its transliteration using a naive rule based phonetic transliteration engine. The news story linking component is depicted in Figure 1.

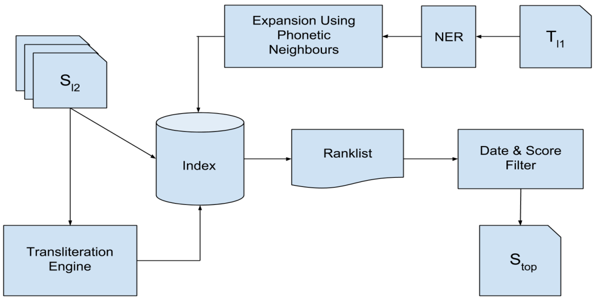

Figure 1: System architecture for news story linking.

Each t_j is queried on the index to find most relevant source document s_i. The query is composed of named entities in the t_j and its transliteration equivalents. We find phonetic neighbours of the query term in transliterated vocabulary space using Editex algorithm (Zobel and Dart, 1996). Such expanded query is used to retrieve source documents from the index with the goal to find the same news story in the language. A date and score based filtering is applied on the ranklist to remove news story which were published beyond a distance of ±n days from t_i and have similarity score less than a threshold. This is inspired by the findings reported in (Barker and Gaizauskas, 2012) that news stories with same news event and focal event tend to be published in a close vicinity of the target news story dates and are expected to contain parallel or comparable content.

3.2. Dictionary Generation

Once news documents are linked across languages, content of news document pairs are cleaned and stop-words are removed from these documents. All the words in the corresponding Hindi document are associated with each word in the English document. These co-occurring words from all pairs of en-hi documents from entire corpus are then added to a map along with their frequency. Most frequent co-occurring word pair is added to our dictionary and then all other word pairs containing either part of most co-occurring word pair are discarded. We repeat the previous step till we run out of entries. We also keep a threshold on minimum co-occurrence of word pair to be added to the dictionary.

[1] As the English and German are written using Roman script, simple normalisation steps and diacritics handling would take care of proper nouns matching. If the languages are written using different scripts, it would require to induce transliteration schemes.

4. Evaluation and Analysis

We evaluate our approach on both the subtasks. The evaluation of news story linking is carried on a benchmark collection of news story linking task. Once the satisfactory performance is achieved and parameters are tuned, we run the dictionary generation module.

4.1. News Story Linking

For the news story linking subtask evaluation, we use publicly available CL!NSS 2012 dataset[2]. The corpus contains around 50k Hindi news stories in the source collection which is the 2010 crawl of Navbharat Times[3]. The target collection contains few selected English articles from Times of India[4]. The task is to link each target news story to source news stories which have the same focal event and news event. More details about the task and the dataset can be found in (Gupta et al., 2012).

Position	Best System	Our Method
1	0.3229	0.3659
2	0.3177	0.378
3	0.3106	0.3581
4	0.3151	0.3854
5	0.3259	0.3879
10	0.338	0.3978
20	0.36	0.4162
50	0.3741	0.4689

Table 1: Performance evaluation of the News story linking module.

Table 1 presents the performance evaluation of the news story linking module and its comparison with the best system of the last year of the shared task. We also plot the distance of the relevant documents from its target news story in terms of publication date which is depicted in Fig 2. It should be noted that most of the relevant news stories are published within the date window of 10 days.

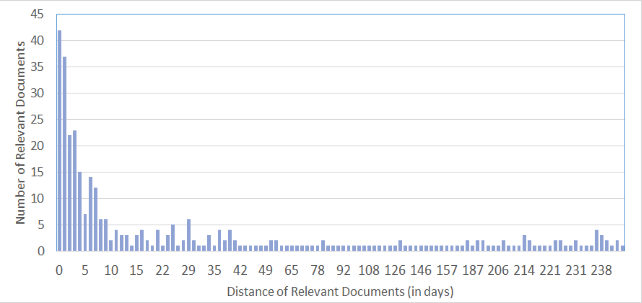

Figure 2: Analysis of relevance judgment data of CL!NSS data.

Once we tune the parameters of our approach on the benchmark collection, we crawl NDTV - a news agency archives[5] for years 2011 and 2012. We also crawled Times of India and Navbharat Times for English and Hindi news stories respectively for 5 years (2008-2012) referred as Times. The statistics of the crawl is depicted in Table 2. We run our algorithm on the NDTV and Times datasets separately to find the news story links among them. The number of links found are also listed in Table 2.

Year	English	Hindi	Pairs
2008	98198	61248	24268
2009	125527	55509	26333
2010	148081	72625	59815
2011	221801	117157	95721
2012	227070	116020	132478
2011-12 (NDTV)	91838	27560	10160

Table 2: Statistics of the news stream data for years 2008-12 for Times of India and Navbharat Times and 2011-12 for NDTV.

4.2. Dictionary generation

We ran the dictionary generation algorithm on the discovered news story links across the languages. The dictionaries generated with our method contains around 2550 mappings for NDTV and around 4000 mappings with overall accuracy of 45%. The dictionaries are manually evaluated. Figure 3 presents the accuracy diagram for the dictionaries from the top position to the last one. It should also be noted that the on top positions, the method is able to find word translations with high quality. The dictionary generated with Times data shows an improvement in quality *e.g.* there are ∼2000 mappings over the accuracy threshold of 0.6 while with NDTV there are ∼1000.

We compare our dictionary with that generated from Wikipedia as described in (Dubey and Varma, 2013). We first check which terms are present in both the dictionary and their accuracy. We also check the dictionary entries found by our method which are not present in the Wikipedia dictionary. Venn diagram in Figure 4 shows two sets of dictionaries and Table 3 presents their statistics.

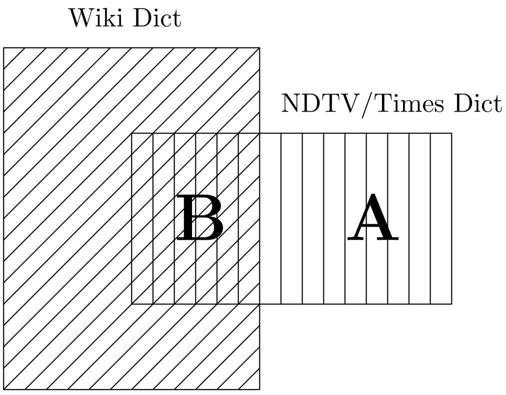

Figure 4: Venn diagram of dictionary coverage for the generated dictionaries (NDTV/Times) and the dictionary generated using Wikipedia.

[2] http://www.dsic.upv.es/~pgupta/clinss/2012/

[3] http://navbharattimes.indiatimes.com/

[4] http://timesofindia.indiatimes.com/

[5] http://archives.ndtv.com/articles/2011-01.html

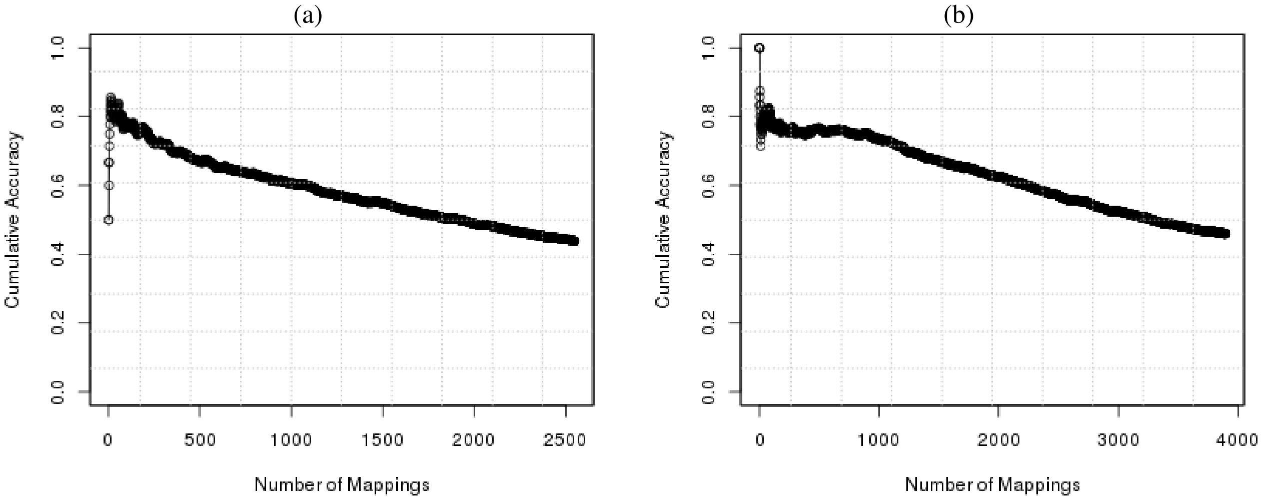

Figure 3: Accuracy analysis of dictionary entries: (a) NDTV and (b) Times

Partition	Size	Accuracy
A (NDTV)	709	36%
A (Times)	2231	31%
B (NDTV)	1841	47%
B (Times)	1662	67%
B (Wiki)	1841	72%

Table 3: Accuracy and coverage statistics of the generated dictionaries and comparison with the dictionary generated using Wikipedia based on Figure 4.

As can be noticed from the Table 3, dictionary generated by Wikipedia is of high quality but it should also be noted that Wikipedia is a comparable corpus where articles are topically aligned and the articles are highly structured *e.g.* titles and sections. On the contrary, our algorithm is able to find word translations from a completely unlabeled data from the web which is growing continuously and with a much higher rate than Wikipedia. It can also be observed that the quality of the dictionary increases with incorporation of more data *e.g.* the dictionary with Times corpus is of higher quality than that of NDTV corpus.

Many of the negative entries in the dictionary generated by our algorithm are very closely topically related. Table 3 presents some of such negative results from the generated dictionary. For example, term “deficit” is associated to (Hindi equivalent of) “economy”, “rebels” with “Libiya”, “ball” with “catch” (in context of cricket), “Agarkar” with “Daredevils” where former is the cricketer of latter team. It also makes our dictionary interesting for some cross-lingual applications like cross-language information retrieval (CLIR) and cross-lingual text categorisation. The generated dictionaries and the code are made publicly available [6].

English Word	Hindi Word
deficit	अर्थव्यवस्था (economy)
rebels	लिबिया (Libiya)
ball	कैच (catch)
Agarkar	डेयरडेविल्स (Daredevils)
degrees	तापमान (temperature)
facebook	जुकरर्बग (zuckerberg)

Table 4: Incorrect but closely related entries in the generated dictionaries.

5. Remarks

In this study we presented a completely automatic method to generate cross-lingual dictionary from quasi-comparable collection to address concrete question: *Can continuous stream data enrich bilingual term translations without any prior knowledge?* Our method is compared with the dictionary generated with a structured and annotated resource like Wikipedia with 30k comparable paired documents. To support our ideas, the dictionary is generated over a small crawl of news media using 10k comparable pairs and large crawl of Hindi and English news stories of around 338k comparable pairs. The main advantage of our method is entries are sorted by its confidence level (c.f. Figure 3), hence top-N partition of the dictionary can be used depending on the application. Term level translations at top positions in our dictionary are highly accurate and such associations grow with the amount of pairs induced in the dictionary generation module. Moreover, our dictionary is able to find terms which are not present in dictionary generated by Wikipedia. On incorporation of more data both content and quality of the dictionary increases. In depth analysis of the generated dictionary reveals that method is able to find some topical associations on top of exact translations. The generated resource and code is made publicly available. Future work should address the issue of polysemy and disambiguation in the evaluation. One should also try to map the multiword units especially named entities during extraction process.

[6] `http://www.dsic.upv.es/grupos/nle/quasi-dict/`

Acknowledgment

This work is carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie. The work of the second and fourth authors is also supported by the DIANA APPLICATIONS Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

6. References

Ahmed, U. Z., Bali, K., Choudhury, M., and VB, S. (2011). Challenges in designing input method editors for indian lan-guages: The role of word-origin and context. In *Proceedings of the WTIM '11*, pages 1–9, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Barker, E. and Gaizauskas, R. J. (2012). Assessing the comparability of news texts. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *LREC*, pages 3996–4003. European Language Resources Association (ELRA).

Braschler, M. and Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *ECDL*, pages 183–197.

Chu, C., Nakazawa, T., and Kurohashi, S. (2013). Accurate parallel fragment extraction from quasi-comparable corpora using alignment model and translation lexicon. In *IJCNLP*, pages 1144–1150.

Dubey, A. and Varma, V. (2013). Generation of bilingual dictionaries using structural properties. *Computación y Sistemas*, 17(2).

Dumais, S., Landauer, T. K., and Littman, M. L. (1997). Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing.

Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING*.

Gupta, P., Clough, P., Rosso, P., and Stevenson, M. (2012). PAN@FIRE: Overview of the cross-language !ndian news story search (CL!NSS) track. In *Fourth International Workshop of Forum for Information Retrieval Evaluation (FIRE)*, pages 1–13.

Gupta, P., Clough, P., Rosso, P., Stevenson, M., and Banchs, R. E. (2013). PAN@FIRE 2013: Overview of the cross-language !ndian news story search (CL!NSS) track. In *Fifth International Workshop of Forum for Information Retrieval Evaluation (FIRE)*.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504, December.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Palkovskii, Y. (2012). Working note for CL!NSS. In FIRE, editor, *FIRE 2012 Working Notes. Fourth International Workshop of the Forum for Information Retrieval Evaluation*.

Platt, J. C., Toutanova, K., and Yih, W.-T. (2010). Translingual document representations from discriminative projections. In *EMNLP*, pages 251–261.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 58–65, New York, NY, USA. ACM.

Zobel, J. and Dart, P. (1996). Phonetic string matching: lessons from information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 166–172, New York, NY, USA. ACM.