International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 3923~3932
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp3923-3932

Two Level Disambiguation Model for Query Translation

Pratibha Bajpai[1], Parul Verma[2], Syed Q. Abbas[3]
[1,2]Department of Information Technology, Amity University, India
[3]Department of Computer Science, Ambalika Institute of Management and Technology, India

Article Info

Article history:

Received Oct 5, 2017
Revised Jan 16, 2018
Accepted Jul 19, 2018

Keyword:

Coherence model
English-hindi cross language information retrieval
Query translation
disambiguation

ABSTRACT

Selection of the most suitable translation among all translation candidates returned by bilingual dictionary has always been quiet challenging task for any cross language query translation. Researchers have frequently tried to use word co-occurrence statistics to determine the most probable translation for user query. Algorithms using such statistics have certain shortcomings, which are focused in this paper. We propose a novel method for ambiguity resolution, named 'two level disambiguation model'. At first level disambiguation, the model properly weighs the importance of translation alternatives of query terms obtained from the dictionary. The importance factor measures the probability of a translation candidate of being selected as the final translation of a query term. This removes the problem of taking binary decision for translation candidates. At second level disambiguation, the model targets the user query as a single concept and deduces the translation of all query terms simultaneously, taking into account the weights of translation alternatives also. This is contrary to previous researches which select translation for each word in source language query independently. The experimental result with English-Hindi cross language information retrieval shows that the proposed two level disambiguation model achieved 79.53% and 83.50% of monolingual translation and 21.11% and 17.36% improvement compared to greedy disambiguation strategies in terms of MAP for short and long queries respectively.

Corresponding Author:

Pratibha Bajpai,
Department of Information Technology,
Amity University,
Lucknow, India.
Email: pratibhabajpai@gmail.com

1. INTRODUCTION

The two commonly used linguistic resources used for query translation are parallel corpora and bilingual dictionaries. Algorithms based on parallel corpora estimate the translation of query words by finding the association between words of the source language and words of the target language. Examples in this category include relevance language models [1]-[3] and statistical translation models [4]-[7]. The major shortcoming of these methods is the availability of parallel bilingual corpora, especially for low resource languages.

Dictionaries, on average offer a good number of translation alternatives for each source query term. The simplest scheme to this problem is to use all alternatives, a method adopted by [8], [9]. This cannot be approved as ambiguity resolution. Other researchers study co-occurrence patterns of query terms in large document collection for sense disambiguation [10]-[13]. Suppose the two terms 'railway' and 'coach' are present in source language query. The term 'coach' has three senses (rail coach; carriage; instructor) in bilingual dictionary. Presence of other term 'railway' in the query, suggest that 'coach' is unrelated to carriage or instructor. Thus it can be rationally predicted that a correct translation of 'coach' will tend to co-occur with translation of 'railway' in target language corpus.

Approaches based on above idea deal with translation ambiguity by computing the coherence score of a translation candidate to the entire query. A translation candidate has a high coherence score if it frequently co-occurs with the translations of other query terms. Finally the translation with highest coherence score is selected for the query term under consideration [1], [14], [15]. In this way these approaches make a binary decision for each translation. This is not functional when we have a number of translations of a query term with similar coherence score. Likewise given the context of query, these approaches do not prioritize the translation alternatives of a query term and treat them equally. This may degrade the effectiveness of any CLIR system. Also the selection of a translation of a query term on the basis of high coherence score is obtained independently from the translations of other query words thereby leading to local solutions.

To overcome these shortcomings of previous works, we propose a novel model, named 'two level disambiguation model'. The model performs disambiguation at two levels. At first level, we call it 'local disambiguation'. Local disambiguation provides a proper distribution of importance factor for translation candidates indicating their relevancy in the given context. This will impact the effectiveness of our CLIR system. At next level we perform 'global disambiguation', which scan all possible permutations of translation candidates to select the best one and then form the target language query by combining its elements .This eradicates the problem of translations being selected independently. It's for the first time that query terms have been disambiguated twice- once in local context and secondly in context of the entire query. Two level disambiguation is advantageous in terms of increasing the number of relevant documents retrieved against the user queries. This has been reported in Section 4 under experimental results.

The rest of the paper is structured as follows: Section 2 briefly reviews the related work in selection-based approaches for query translation disambiguation. Section 3 describes our two level disambiguation model, along with an example to demonstrate the working of proposed model. Section 4 presents the experimental results. Section 5 concludes this work.

2. RELATED WORK

The effectiveness of a dictionary based query translation depends highly on its competence in resolving ambiguity [15], [16]. To find the correct translation of a query term, researchers have tried exploiting the context of query in terms of co-occurrence statistics. Co-occurrence statistics emphasizes that the correct translations of individual query terms tend to co-occur in the target language corpus while incorrect translations do not. The good translation word is the one which has high coherence with the translations of other query words and is hence selected as the correct translation of the source query term.

Ideally, the selection of a translation of a query term should depend only on the selected translations of other query terms. But to lower the computation cost, previous works using coherence model proposed an approximate greedy algorithm to select the best translation alternative, including both selected and unselected translations for all query terms. The approximate greedy algorithm is stated as follows:

Greedy algorithm for disambiguation of translation candidates of query terms

1. Source query is represented as a set {(e_1, H1), (e_2, H2),..... (e_n, H_n)}, where e_i is the source query term and H_i=(h_{i1},h_{i2}…..h_{ij}) is the list of translation candidates of e_i obtained from bilingual dictionary.
2. For each H_i,
   2.1. For each translation h_{ij}∈H_i, define the similarity measurement between the translation h_{ij} and a set H_k($k\neq i$). Cohesion of h_{ij} with respect to H_k is the maximum similarity of h_{ij} with every $h_{kl} \in H_k$. So,

   $$\text{sim}(h_{ij}, H_k) = \boldsymbol{argmax}_{\boldsymbol{h_{kl}\in H_k}, \boldsymbol{k\neq i}} \boldsymbol{sim}(\mathbf{h}_{ij}, \mathbf{h}_{kl}) \quad (1)$$

   2.2. Compute coherence score for h_{ij} as

   $$\text{Score}(h_{ij}) = \sum_{1\leq k\leq n,\ k\neq i} \text{Cohesion}(h_{ij}, H_k) \quad (2)$$

3. Select the translation h∈H_i with the highest Score.

The set of selected terms h from each H_i, $1\leq i \leq n$ forms the final translated query.

Similarity between the terms can be measured using either dice coefficient [10] or mutual information [13], [17] or its variants [16], [18]. Basically, the best sense for each term is chosen resulting in the final set of selected translations containing translations that are closely related with one another in the context of source query.

Many researchers have used greedy algorithm to disambiguate source language queries. Croft and Ballesteros experimented with Spanish-English language pair to select the translation with the highest coherence score and revealed that the method is very successful for language pairs with scarce resources [15]. Adriani approached the similar problem and used maximum similarity score between translation candidates for different query terms [10]. Later Gao*et al*. claimed that increase in distance between two terms weakens the association between them. They refined the disambiguation algorithm by incorporating decaying factor with the mutual information statistics. This refinement easily outperformed the basic co-occurrence model [18].

Maeda *et al*. revisited the problem in a slightly different manner and instead of considering the co-occurrence of consecutive terms they considered all pairs of possible translations of query terms [13]. Monz and Dorr, determined the solution by an iterative procedure, which is sensitive to the initialization of parameters or the stop criterion employed in the iterative procedure [19]. Zhou *et al*. viewed the co-occurrence of possible translation terms within a given corpus as a graph and determined the importance of a translation using global information recursively drawn from the entire graph [20]. Giang *et al*. used mutual summary score based on word distribution in document collection to outperform basic model [12]. Andres Duque *et al*. technique combines both the dictionary and co-occurrence graph to select the most suitable translation from the dictionary and thereby disambiguating the query. The method relies on the hypothesis that words appearing in the same document tend to share related senses and thereby represent a coherent content. The co-occurrence graph is obtained by considering only those words that frequently co-occur in the same documents. They then use various algorithms to combine information from the two sources [21].

The greedy algorithm selects the best translation of individual query terms considering both selected and unselected translations of other query terms, thereby leading to translations being selected independently. Furthermore, the translation having maximum coherence is only selected as the final translation disregarding other translation alternatives of a query word. This binary decision is not acceptable where translation candidates have similar coherence scores.

3. PROPOSED METHOD

In this section we propose a relatively simple yet effective novel model named "Two level disambiguation model" to address the anomalies of existing approaches. Cross Lingual word sense disambiguation performs disambiguation of source language words while translating them to target language [22]. Consider a source query Q containing say, three terms s_1, s_2 and s_3. Let the target language translations for these terms be $t_{1,1}$; $t_{2,1}$, $t_{2,2}$ and $t_{2,3}$; and $t_{3,1}$and $t_{3,2}$ for s_1, s_2 and s_3 respectively.

In Figure 1 each link between two translation candidates represents co-occurrence frequency of that pair of translation alternative. Co-occurrence frequency between translations of same query term is not considered, thereby leading to no links between them.

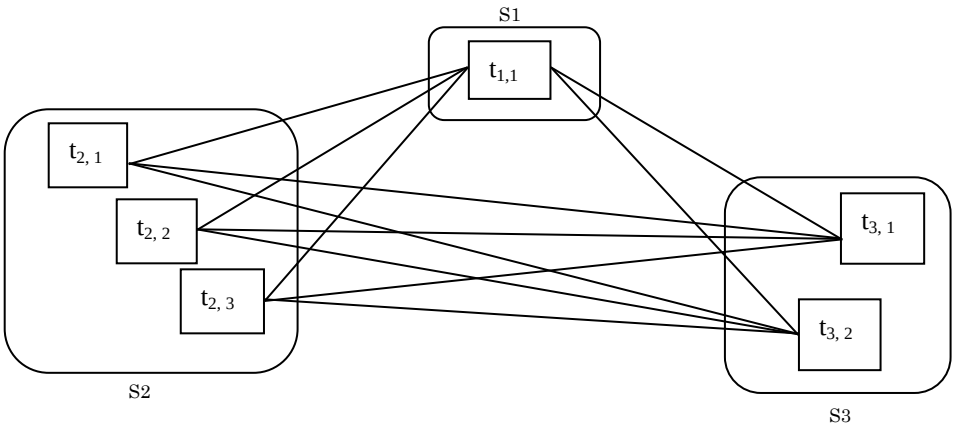

Figure 1. Co-occurrence graph for source query Q

Suppose that $t_{2,1}$ occurs more frequently with $t_{3,2}$ than any other pair of translation candidates for s_2 and s_3. As a result $t_{2,1}$ and $t_{3,2}$ would be more ideal translations for s_2 and s_3. On the other hand, let's assume that $t_{2,1}$ and $t_{3,2}$ do not co-occur with $t_{1,1}$ at all, but $t_{2,2}$ and $t_{3,1}$ do. This raises a very valid question

as to select which pair of translation candidates i.e. (1) $t_{2,1}$ and $t_{3,2}$ or (2) $t_{2,2}$ and $t_{3,1}$. To have a better understanding, consider the English source query-- "*Security measures in railway coach"*
Dictionary translations for query terms are:

Security = {सुरक्षा, जमानत}

Measure = {उपाय, राशि , मापदण्ड}

Rail = {रेल}

Coach = {कोच, प्रशिक्षक }

In the context of 'rail', the pair { रेल, राशि}will be preferred over the pair { रेल, उपाय} and { रेल , मापदण्ड } but if we talk about the 'security of railway coach' the combination { रेल , सुरक्षा ,उपाय }is ideal than the combination { रेल, सुरक्षा , राशि }.This implies that disambiguation at local level only is not ideal, but the need to perform disambiguation globally considering the query as a single concept is required too. To address these anomalies of existing approaches, we propose a relatively simple yet effective novel model named "Two level disambiguation model" which performs disambiguation at two levels: First level disambiguation and Second level disambiguation.

3.1. First level disambiguation

We refer first level disambiguation as 'local disambiguation'. First level disambiguation deal with the translation candidates in pairs only. This is done with the aim to obtain partial data for the likelihood of a translation in the perspective of other query terms. For a given query word, instead of taking binary decision for its translation candidates, we calculate the importance factor of each of the candidates in the context of given query. This importance factor approximates the probability of a candidate to be selected as a final translation of a query word. Higher the importance factors more it is relevant in the context of the user query. A translation candidate is assigned a high importance factor if it is rational with the semantic meaning of the user query.

Let the source query be Q = $\{q_1, q_2, \dots \dots . q_n\}$

Step 1

a) Find the translation candidates from bilingual dictionary. Let the translation candidates of query term q_i be represented as a set $T_i = \{t_{i1}, t_{i2}, \dots . . t_{im}\}$.
b) For each t_{ij}, where 1<=i<=n and 1<=j<=m retrieve all example sentences for its synset, hypernyms and homonyms from Hindi WordNet. Example sentences from other sources are also added for t_{ij}. Store them in a file.

Step 2

a) Assign a 2*2 usage matrix M_i for each query term q_i. The columns represent the translation candidates of query term q_i while rows represent the translation candidates of remaining query words q_k, where 1<=k<=n and k≠i. Initialize the matrix with 0's.
b) Count the usage of a translation candidate t_{ij} of q_i in example sentences of translation candidates t_{kl} of other query terms q_k, where 1<=k<=n and k≠i. The count is stored in matrix M_i .
c) Repeat the same for all translation candidates of all query terms.
d) Find the sum of column entries to obtain UC_{ij} ,the Usage Count of a particular translation candidate with respect to translation candidates of other query terms.
e) Normalize UC_{ij} to obtain IF_{ij}, Importance factor of translation candidate t_{ij}.

3.2. Second level disambiguation

We refer second level disambiguation as 'global disambiguation'. Global disambiguation aims at finding the most suitable translation for the given query. This resolves the problem of translations being selected independently from selected and unselected translations of remaining query terms. This step computes the coherence between all possible combinations of translation candidates of query terms. To give due regard to most preferred translation candidates , the algorithm combines dice coefficient score with the importance factor for word pairs to obtain Weighted Summary Dice Coefficient (WSDC) for every combination obtained by including one translation candidate for each source query term q_i. The motivation behind using Dice coefficient for measuring association strength between two terms is that the value of the Dice coefficient ranges between 0 and 1 (where 1 is perfect co-occurrence), whereas mutual information has no upper bound [19].

Step 3

a) Find all combinations C= $\{h_1, h_2, \dots \dots . , h_n\}$ where h_i is a translation candidate of q_i.

b) Compute WSDC for each combination C as

$$WSDC(C) = \sum_{h_i,h_j \in C}(DC(h_i,h_j) * IF(h_i) * IF(h_j)) \quad \text{where } 1<=i<=n,\ 1<=j<=n \text{ and } i \neq j \quad (3)$$

and,

$$\text{Dice Coefficient, } DC(h_i,h_j) = \frac{2*freq(h_i,h_j)}{freq(h_i)*freq(h_j)} \quad (4)$$

$freq(h_i)$= the number of occurrences of term h_i in training corpus
$freq(h_j)$= the number of occurrences of term h_j in training corpus
$freq(h_i,h_j)$= co-occurrence frequency of terms h_i and h_j in a sentence in documents.

c) Select the combination with highest WSDC score as the target language query Q^t of the source query Q.

$$Q^t = arg\ max_C WSDC(C) \quad (5)$$

3.3. Example of disambiguation using proposed model

Reconsider the English source query "Security measures in railway coach". Performing first level disambiguation we obtain IF_{ij}, Importance factor of translation candidate t_{ij} as follows. Table 1 represents the result of first level disambiguation of the proposed model. The result suggest translation set {सुरक्षा, राशि, रेल, कोच} as the most appropriate translation of given English query, depending upon the highest value obtained by the translation candidate of respective query terms.

Table 1. Importance Factor of Translation Candidates Estimated using First Level Disambiguation

S.No.	Source query term	Translation Candidates	Importance Factor
1	Security	सुरक्षा	0.835
2		जमानत	0.164
3	Coach	कोच	0.666
4		प्रशिक्षक	0.444
5	Measure	राशि	0.527
6		उपाय	0.444
7		मापदण्ड	0.027
8	Rail	रेल	1.0

3.4. After second level disambiguation

Figure 2 and Figure 3 represent computation of Weighted Summary Dice Coefficient (WSDC) for translation sets {सुरक्षा, राशि, रेल, कोच} and {सुरक्षा, उपाय, रेल, कोच}. WSDC {सुरक्षा, राशि, रेल, कोच} and WSDC {सुरक्षा, उपाय, रेल, कोच} are 0.651 and 0.713 respectively. Depending upon Weighted Summary Dice Coefficient, after second level disambiguation, translation set {सुरक्षा, उपाय, रेल, कोच} is selected as the final translation for the given example query "Security measures in railway coach". This is because in first level disambiguation, translation candidates are considered in pairs while in second level disambiguation the two translations 'राशि' and 'उपाय' when treated in the context of entire query, 'उपाय' turns out to be correct translation for English query term 'measure'.

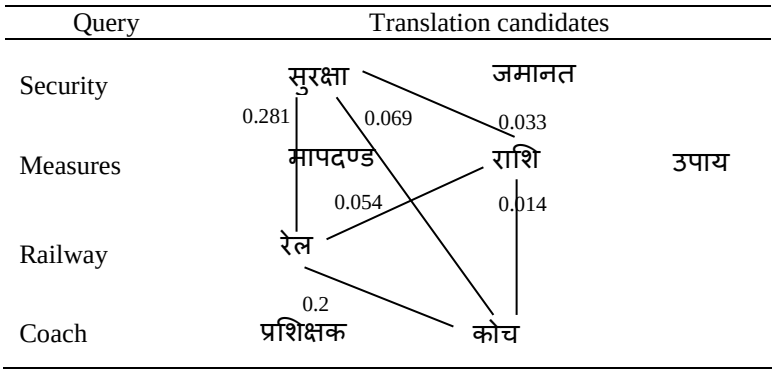

Figure 2. Computation of WSDC for translation set {**सुरक्षा**, **राशि**, **रेल**, **कोच**}

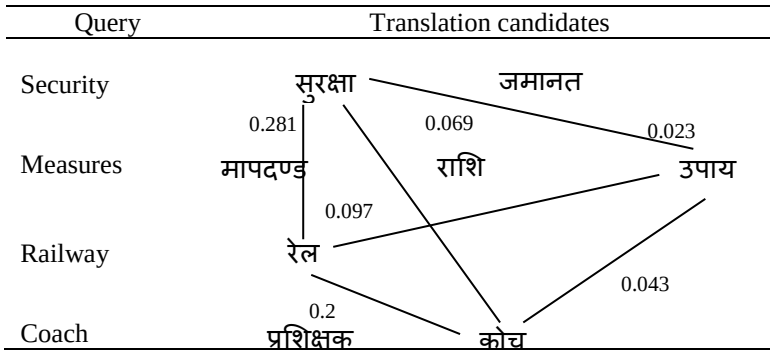

Figure 3. Computation of WSDC for translation set {**सुरक्षा**, **उपाय**, **रेल**, **कोच**}

4. EXPERIMENT

In this section we discuss our evaluation of the 'two level disambiguation' model described above. We first present the set-up of our experiment followed by the experimental results.

4.1. Experimental setting

For training our model, we developed a Hindi corpus that contains 5000 articles in UTF-8 encoding, published in leading Hindi newspapers Dainik Jagran, Amar Ujala and Web Dunia with an average size of 25 KB each. The document collection consists of articles across the domains such as politics, sports, science, entertainment, social science and criminal, motivated by the heterogeneous nature of user queries. We create a set of 50 English queries as per the CLEF & TREC guidelines to evaluate the performance of the proposed model. The test queries are able to capture the nature of the query posed by web user. We use publicly available online bilingual English to Hindi dictionary 'Shabdanjali' developed in IIIT, Hyderabad and containing 28K Hindi words to translate English queries to Hindi language queries [23]. The dictionary required conversion from ISCII to UTF-8 encoding and some basic normalization. We use an English stop word list of 507 English words to remove stop words from the queries formulated for evaluation. Porter stemming algorithm is used to reduce inflected English query words to base form [24]. Hindi WordNet provided by the Linguistic Data Consortium is a lexical database for Hindi and developed by IIT Bombay [25]. It is used for first level disambiguation. It contains 103438 unique Hindi words and 39271 number of synset. We use it to fetch example sentences for all the senses, hypernyms and homonyms of a translation candidate.

The proposed model is evaluated at actual web documents using Google indexed database. Web search engines contain huge volume of documents covering varied domains and periodically update their index. Thus the set of documents retrieved for each disambiguated query can give good judgment of the efficiency of proposed two level disambiguation model. The relevance judgments for the Hindi documents obtained with respect to English queries is established with the help of three Hindi speaking volunteers from Indian Institute of Technology (Banaras Hindu University). Document which is judged as relevant by all the

three volunteers is marked as relevant else treated as irrelevant. In this way we collected the set of relevant Hindi documents for each English test query.

4.2. Evaluation method

The following methods are compared to investigate the effectiveness of our model for query translation and disambiguation:

a. Monolingual: retrieval using the Hindi queries translated manually by Hindi language expert. Monolingual run provides unreachable performance ceiling for any cross lingual information system as translation process is inherently noisy.
b. Simple translation: retrieval using query translation by taking the first translation from the bilingual dictionary. The first translation for any term in bilingual dictionary is generally the most frequent translation for that term according to World Wide Web.
c. Base approach: retrieval using basic Greedy algorithm to find best translation, as described in section II. We use the same training document collection to estimate cohesion scores, which is prepared to train our model.
d. Proposed model: retrieval using the proposed two level disambiguation model.

4.3. Experimental results

The test query set consists of two types of queries. The first is termed as short queries and the other as long queries. Short query comprises of 2 to 4 keywords whereas long query is formed as natural sentence with average length of 7.12 terms. Short queries are the actual representation of most queries posed by users, particularly the web queries which tend to have few terms. Thus we chose to have major number of queries in our test query set as short queries.

We have used standard evaluation measure, Mean Average Precision (MAP) to evaluate our proposed model with monolingual, simple and base approach. The evaluation has been done on first 50 Hindi documents retrieved using Google search engine. Table 2 describes our experimental results. For each method, we give average values of P@k with k= 10, 20, and 50.

Table 2. Run Statistics for short Queries

Experimental Run	Mean Average Precision (MAP)	Percentage Monolingual
Monolingual	0.518	--
Simple translation	0.200	38.61%
Base Approach	0.325	62.74%
Two level Disambiguation	0.412	79.53%

Table 3 compares the MAP value of simple translation, base approach and proposed method with baseline method i.e. monolingual run for short queries. The performance of these runs is 38.61%, 62.74% and 79.53% respectively of monolingual run. The proposed approach shows an improvement of 21.11% over the base approach.

Table 3. Average Retrieval Precision of Experimental Runs for Short Queries

Experimental Run	P@10	P@20	P@50
Monolingual	0.483	0.420	0.309
Simple translation	0.145	0.112	0.089
Base Approach	0.316	0.270	0.184
Two level Disambiguation	0.383	0.336	0.240

Table 4. Average retrieval precision of experimental runs for long queries

Experimental Run	Mean Average Precision (MAP)	Percentage Monolingual
Monolingual	0.600	--
Simple translation	0.263	43.83%
Base Approach	0.414	69.00%
Two level Disambiguation	0.501	83.50%

Table 4 compares the MAP value of simple translation, base approach and proposed method with baseline method i.e. monolingual run for long queries. The performance of these runs is 43.83%, 69.0%and 83.50% respectively of monolingual run. The proposed approach shows an improvement of 17.36% over the base approach.

4.4. Analysis

The greedy approach used to disambiguate query words treats all translation alternatives equally. But there exists a significant variance in the priority across different Hindi words, as demonstrated in Figure 4.

	जीवन	आयु	कार्यकाल	जोश
Life	*0.249*	*0.247*	*0.248*	*0.256*
	हमला	वार	दौरा	चढाई
Attack	*0.262*	*0.23*	*0.337*	*0.171*
	उपाय	राशि	मापदण्ड	
Measure	*0.444*	*0.527*	*0.027*	

Figure 4. Examples of Importance factor estimated by first level disambiguation

The first example in the Figure 4 shows an almost uniform distribution over all translation alternatives, while the third one is a skewed distribution. In between, the second example is a case which is neither uniform nor skewed. These three examples illustrate why we measure the importance of each of the candidates in the context of given query.

The base approach which also exploits word co-occurrence statistics for query translation disambiguation shows a performance drop of 21.11% over the proposed approach. Consider a query "Security measure in railway coach". The base approach makes incorrect translation selection for the term 'measure' as 'राशि'. The correct Hindi translation is 'उपाय' instead of 'राशि'. This is because greedy algorithms do not consider the query as a single concept and disambiguate the query terms independently in pairs. The translation candidate 'राशि' for term 'measure' is more consistent with either of the query terms 'रेल', 'कोच' and 'सुरक्षा'as compared to translation alternative 'उपाय', thereby leading it to be selected as the final Hindi translation for 'measure'.

The proposed approach achieves 79.53% of monolingual run in terms of MAP. The reason behind it is the treatment of some words by the dictionary used for bilingual translation of source query words. For instance, for source query "Indian animation industry films", the term 'animation' is translated as 'उत्साह', 'जीवंतता', 'जीव-संचारण', 'जीवंतता' etc. by dictionary. These translations provided by the translation dictionary are inappropriate in the given context. The documents retrieved against these translations describe journey of Indian film industry instead of role of animation industry in Indian cinema.

The simple translation run shows the worst performance among all the runs. In simple translation we take the first translation from the bilingual dictionary for each query term. The first translation for any term in bilingual dictionary is generally the most frequent translation for that term according to World Wide Web. The context of the query is not exploited at all for disambiguation and thereby leading to maximum degradation in performance as compared to monolingual run. To fully examine the effectiveness of our proposed model, we test it against both the long English queries and the short English queries. The results show that the use of the proposed query translation scheme is more effective with longer queries than with shorter queries. This is expected because longer queries provide multiple contextual words which can contribute to better disambiguation. This result confirms our intuitive assumption that natural sentence based queries are less ambiguous than keyword based queries. Proposed approach does not show much significant improvement over base approach for longer queries. This is convinced as both approaches depend on context of query for disambiguation. Rich context of long queries help both approaches in successful disambiguation

of source query words.

Figure 5 shows the MAP score comparison of the four experimental runs for both short queries and long queries.Our approach can easily be implemented for other pair of Indian languages. The approach is simple and uses only a lexical database, bilingual dictionary and a monolingual corpus for query translation and disambiguation. However the success rate for other languages may vary due to the unavailability of resources in a particular language.

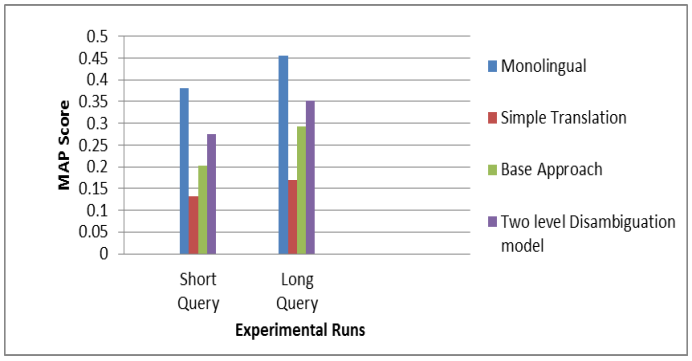

Figure 5. MAP score comparison of various experimental runs

5. CONCLUSION

In this paper, we propose a new model for cross language information retrieval system, named "two level disambiguation model". Compared to previous selection based approaches, the merits of our model are (a) proper distribution of importance factor for translation candidates which indicates their relevancy in the given context, (b) estimation of translations of all query words simultaneously. The results demonstrate effective retrieval by achieving 79.53% for short queries and 83.50% for long queries of the monolingual result. The proposed model shows an improvement of about 20% over the base approach. The results also confirm the general pattern that disambiguation of long natural language sentence query is more effective than short queries. Our method can easily be extended to other language pairs.

The proposed model for cross language information retrieval relies heavily on the coverage of the dictionary and the quality of lexicon used. So, we plan to work on other generic approaches for query translation and disambiguation like using web etc in future.

REFERENCES

[1] Kraaij W. R., *et al.,* "Twenty-one at TREC-8: Using Language Technology for Information Retrieval", in *E. M. Voorhees and D. K. Harman, editors, "The Eigth Text Retrieval Conference (TREC-8),"* National Institute of Standards and Technology, NIST, 2000. NIST Special Publication 500-246, vol. 8, pp. 285-300, 2000.

[2] Lavrenko V. and Croft W. B., "Relevance based Language Models", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM Press, New York, pp. 120-127, 2001.

[3] Lavrenko V., *et al.,* "Cross-lingual Relevance Models", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM Press, New York, pp. 175-182.

[4] Xu J. and Weischedel R., "TREC-9 Cross-Lingual Retrieval at BBN", *The 9th Text Retrieval Conference (TREC-9*), 2002.

[5] Federico M., and Bertoldi N., "Statistical Cross-language Information Retrieval using Nbest Query Translations", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM Press, New York, pp. 167-174, 2002.

[6] Nie J. Y. and Simard M., "Using Statistical Translation Models for Bilingual ir", *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF '01,* Springer-Verlag, New York, pp. 137-150, 2002.

[7] Kraaij W., *et al.,* "Embedding Web-based Statistical Translation Models in Cross-language Information Retrieval", *Comput. Linguist. 29,* vol. 3, pp. 381-419, 2003.

[8] Daelemans W., *et al.,* "Different Approaches to Cross Language Information Retrieval", number 37 in Language and Computers: Studies in Practical Linguistics, Amsterdam, Rodopi, 2001.

[9] Davis M. W., "New experiments in cross-language text retrieval at NMSU"s computing research lab," *The 5th Text Retrieval Conference (TREC-5), D. K. Harman, Ed. NIST, Boulder, CO,* 1996.

[10] Adriani M., "Using Statistical Term Similarity for Sense Disambiguation in cross-language Information Retrieval", *Inf. Retr. 2*, vol. 1, pp. 71-82, 2000.
[11] K. W. Church and P. Hanks, "Word Association Norms Mutual Information and Lexicography", *Computational Linguistics*, vol. 16, no. 1, pp. 23-29, 1990.
[12] Giang L. T., *et al.*, "Experiments with Query Translation And Reranking Methods", Vietnamese-English Bilingual Information Retrieval. SOICT'13, Danang, Vietnam, 2013.
[13] Maeda A., *et al.*, "Query term Disambiguation for Web Cross-language Information Retrieval using a Search Engine", *Proceedings of the 5thInternational Workshop on Information Retrieval with Asian Languages (IRAL'00)*, ACM Press, New York, pp. 25-32, 2000.
[14] Hull D. A. and Grefenstette D. A., "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval", *Proceedings of the 19th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp.49-57, 1996.
[15] Ballesteros L. and Croft W. B., "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval", *Proceedings of the 20th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 84-91, 1997.
[16] Gao J., *et al.*, "Improving Query Translation for Cross-language Information Retrieval using Statistical Models", *Proceedings of the 24th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 96-104, 2001.
[17] Jang M. G., *et al*., "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting", *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
[18] Gao J., *et al.*, "Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 183-190, 2002.
[19] Monz C. and Dorr B., "Iterative Translation Disambiguation for Cross-language Information Retrieval", *Proceedings of the 28th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pp. 520-527, 2005.
[20] Zhou D., *et al*., "Disambiguation and Unknown Term Translation", *Cross Language Information Retrieval*, Springer-Verlag Berlin Heidelberg (CLEF 2007), pp. 64-71, 2008.
[21] Duque A., *et al.*, "CO-graph: A New Graph-based Technique for Cross-lingual Word Sense Disambiguation", *Natural Language Engineering*, vol. 21, no. 5, pp. 743-772, 2015.
[22] Rekabsaz N., *et al.*, "Addressing Cross-lingual Word Sense Disambiguation on Low-Density Languages: Application to Persian", 2017; arXiv.org > cs > arXiv:1711.06196.
[23] Shabdanjali English-Hindi Dictionary from IIIT Hyderabad http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html
[24] Porter stemmer at : https://www.drupal.org/project/porterstemmer
[25] Hindi WORDNET at: www.cfilt.iitb.ac.in/wordnet/webhwn/

BIOGRAPHIES OF AUTHORS

Pratibha Bajpai. Completed M.Sc (CS) from University of Allahabad in 2003 and M.Tech (IT) in 2011. Presently pursuing P.hd in Computer Science from Amity University, Lucknow, India. My research area is Cross Language Information Retrieval for Indian languages.

Dr. Parul Verma. Assitant Professor in Amity University, Lucknow. Completed her P.hd in Computer Science from Ambedkar University, Lucknow in 2012. Her area of research are Sense Disambiguation, Semantic Web, Information Retrieval, Ontologies etc.

Prof. (Dr.) Syed Qamar Abbas. Currently working as Director General, Ambalika Institute of Management & Technology, Lucknow. He has completed M.S. (Computer Science) from BITS PILANI. He has been awarded Ph.D in "Computer Oriented study of Queuing models". He has 24 years of teaching experience and has supervised 15 Ph.D. thesis. He has 90 publications to his credit.