ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Improved Query Translation for English to Hindi Cross Language Information Retrieval

Pratibha Bajpai[1], Parul Verma[2], S.Q.Abbas[3]

Research Scholar, Department of Information Technology, Amity University, Lucknow, India[1]

Assistant Professor, Department of Information Technology, Amity University, Lucknow, India[2]

Professor, Department of Computer Science, Ambalika Institute of Management and Technology, Lucknow, India[3]

ABSTRACT: Bilingual dictionaries have always been an important source of query translation in Cross Language Information Retrieval. Besides other issues bilingual translation suffers from ambiguity problem. To resolve this issue, several recent works have recommended the use of term co occurrence statistics. Same concept with a major modification is the focus of our work described here. Our work is based on the fact that all terms do not have same discriminating power in a query. To overcome such problem, our algorithm provides more weight to discriminating terms in the query and treats co occurrences of useful terms as more valuable than those of frequent terms. The paper also takes into account the concept of local context in formulating formula for co-occurrences statistics. In the experiments, our method achieved 85% of monolingual translation in terms of the mean average precision (MAP). The results are quiet encouraging as compared to other methods used for cross language information retrieval for Indian languages.

KEYWORDS: Cross Language information retrieval; Query Translation; Disambiguation; Co-occurrence

I. INTRODUCTION

The traditional monolingual Information System (IR) facilitates users to access document written in the same language as the query. With the enormous increase of information in various languages on the web, retrieval engines are forced to cross the language barrier and allow users to search for information resources in languages other than the language of the query submitted. This trait of retrieval engines is termed as Cross Language Information Retrieval (CLIR). Cross language Information Retrieval can thus be defined as retrieving documents in language different from the language of request [1].

Achieving effective CLIR is an interesting challenge for researchers. To resolve language disparity, either query or documents can be translated [2]. Although, high quality machine translation system makes it possible to translate documents [3], [4], query translation is more popular in research community. This is because of the shorter length of queries as compared to documents, which make query translation simple and economical.

For query translation, one can use machine translation service or train a system using parallel corpora or employ easy available online machine readable dictionaries (MRDs) [5], [6], [7], [8], [9]. Easy availability of machine readable bilingual dictionaries has made them a viable source for Cross lingual query translation. Each lookup in the dictionary gives back a number of translations of a query word. For instance, word ‘bank’ has three senses. Different senses refer to a financial institution or river bank or reservoir. This is referred as ambiguity problem. Selecting the most appropriate translation from this pool, termed as disambiguation is a crucial part of dictionary translation.

Most of these disambiguation strategies exploit word co-occurrence patterns [10], [11], [12], [13]. 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. This data is quiet helpful as we like to choose the best translation of the query term under consideration that is consistent with the translations selected for all remaining query terms [14].

Gao et al. mentioned that effectiveness of cross lingual query translation is less than 60% as compared with monolingual retrieval in terms of average precision [15].

To improve the average precision, we are implementing a cross lingual English-Hindi retrieval system and check whether we can overcome the mentioned average precision limitation. Our system introduces a method called

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Weighted Mutual Information Score which provides more weightage to the discriminating terms while finding the co-occurrence of query terms.

The paper is structured as follows. Section 2 provides overview of few works related to the use of co-occurrence information to deal with the problem of translation ambiguity. Section 3 discusses our proposed method and disambiguation algorithm and section 4 provides our test results. Finally section 5 concludes our study and gives an outlook on future work.

II. RELATED WORK

Many researchers favoured to use bilingual dictionaries for query term translation as the approach being simple and practical. But the method suffers from the problem of translation ambiguity as there is often one-to-many translation in bilingual dictionaries. So to achieve high performance dictionary based query translation, researcher's resolved ambiguities by making use of Mutual Information statistics [11] to measure frequency of co-occurrence of query terms in existing corpora.

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 [16].

Adrani 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 refined easily outperformed the basic co-occurrence model [17].

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]. In the same year Liu et al. published an algorithm on maximum coherence model. They maximized the overall coherence of the query to estimate the translation probabilities of query terms using an iterative machine learning approach based on expectation maximization [18]. Zhou et al. Viewed the co-occurrence of possible translation terms within a given corpus as a graph and determines the importance of a translation using global information recursively drawn from the entire graph [19]. 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. 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 [20].

III. PROPOSED ALGORITHM

Basic co-occurrence statistics aims at selecting correct translation of query terms. But it does not gives due importance to the discriminating terms in the query. Such terms help in improving the precision of the retrieved documents and thus prove themselves to be more useful in query.

Consider the query "Indian government policies against Pakistan terrorism". Here terms like "Indian government" and "Pakistan terrorism" are more dominant than the specific term "policies" in the query. So most of the documents retrieved against this query will describe Pakistan terrorism rather than this specific query. To overcome such problem, our algorithm provides more weight to discriminating terms in the query and treats co occurrences of useful terms as more valuable than those of frequently co-occurring terms.

The usefulness of a query term to retrieve relevant documents can be measured using the standard tf-idf score. The term weight, w_x^i of term *x* in document *i* is computed using the standard tf*idf weighting formula [21] as follows:

$$w_x^i = tf_x^i . \mathrm{idf}_x \quad (1)$$

where idf_x, the inverse document frequency, is computed as follows:

$$\mathrm{idf}_x = \log(n/\mathrm{df}_x) \quad (2)$$

where,

tf_x^i = the number of occurrences of term *x* in document *i*

df_x= the number of documents containing term x in the collection.
n = number of documents in the collection.

So the usefulness of a term in a collection can be given by:

$$w_x^n = \sum_{i=1}^{n} tf_x^i . \text{idf}_x \quad (3)$$

After weight is assigned to query terms, to make term weights scalable, they are normalized as follows:

$$W_x^n = w_x^n / \sum_{y \in C} w_y^n \quad (4)$$

where C is the context of query as described later.

To obtain the association strength between terms, we use a term association measure called Dice coefficient. Thus Weighted Term Similarity (WTS) is given as

$$\text{WTS(t,x)} = \frac{2 * \text{freq(t,x)}}{\text{freq(t)} + \text{freq(x)}} * \frac{w_x^n}{\log(\text{dist(t,x)})} \quad (5)$$

where,
freq(t)= the number of occurrences of term t in corpus
freq(x)= the number of occurrences of term x in corpus
freq(t,x)= co-occurrence frequency of terms t and x in a sentence.

Moreover the dependency of two terms depends upon their distance from each other. Farther the two terms are, weaker is the relationship between them. So we add a distance factor in our calculation of term similarity.

While translating a word ‘t’ (target word), the remaining words(or their translations) in the query form a context ‘C’ that helps determine the correct translation for the target word. For instance, consider the query ‘Security measures in railway coach’. We use bilingual English to Hindi dictionary ‘Shabdanjali’ to find Hindi translations of English query terms. Here if we consider ‘coach’ (with Hindi translations कोच and प्रशिक्षक) as target term then security (सुरक्षा, जमानत), measure (उपाय, राशि, मापदण्ड) and railway (रेल) form the context. Figure 1 illustrates the proposed method.

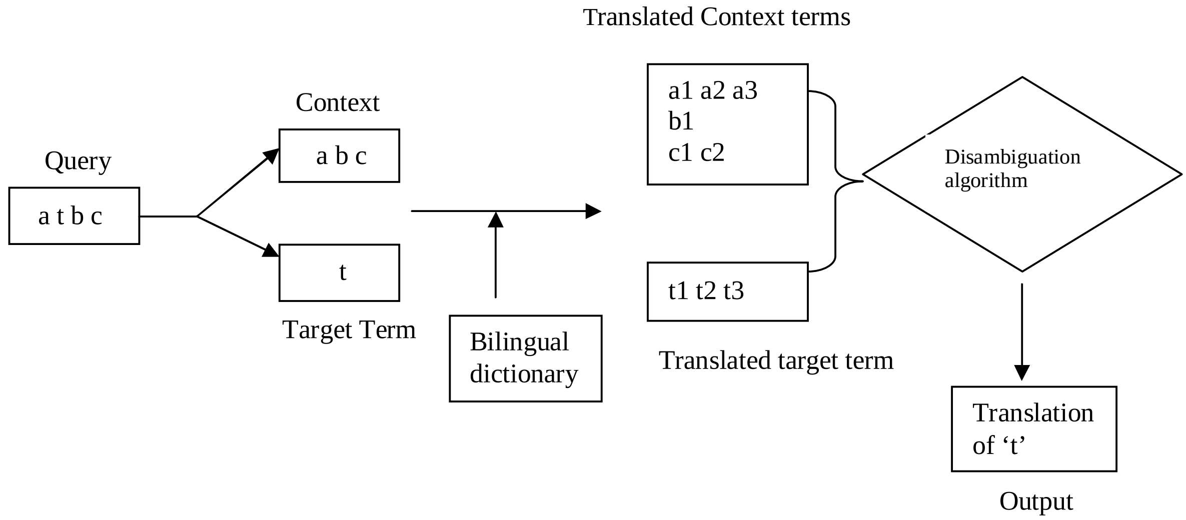


Fig 1: Disambiguation process.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

In Fig. 1 each query term will be treated as a target term one by one considering rest of the query as context. Then target and as well as context terms are translated using bilingual dictionary. These translations are then disambiguated using proposed disambiguation algorithm to obtain suitable translation of target term't'.

A. *Disambiguation Algorithm:*

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

$$\text{Cohesion } (h_{ij}, H_k) = \boldsymbol{argmax}_{h_{kl} \in H_k \ , k \neq i} \boldsymbol{WTS}(\mathbf{h}_{ij}, \mathbf{h}_{kl}) \qquad (6)$$

   2.2. Compute final score for h_{ij} as

$$\mathbf{Score}\ (\mathbf{h_{ij}}) = \sum_{1 \le k \le n,\ k \neq i} \boldsymbol{Cohesion}(h_{ij}, H_k) \qquad (7)$$

3. Select the translation $h \in H_i$ with the highest Score.

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

IV. Results and Discussion

To evaluate the effectiveness of our proposed disambiguation algorithm, we create a test environment having a set of 30 English queries developed on the lines of CLEF queries. The queries are formulated by using only title field of the English topics. We built Hindi corpus consisting of 6000 news articles published in Jagran, a news magazine in Hindi. We use publicly available online bilingual English to Hindi dictionary 'Shabdanjali' developed in IIIT, Hyderabad to translate English queries to Hindi language queries. The dictionary required conversion from ISCII to UTF-8 encoding and some basic normalization.

Table 1: Test Result

Method	P(5)	P(10)	P(15)	MAP	Perf.
Proposed Disambiguation algorithm	.52	.3	.17	.53	85 %
Monolingual	.64	.3	.12	.62	--

The table 1 describes our test results. The precision p(k), for each query returns the fraction from top k documents retrieved from IR system that are relevant. The average precision (AP) is calculated using standard formula

$$\text{AP} = \frac{\sum_{k=1}^{n} p(k) * rel(k)}{N} \qquad (8)$$

where n is the number of retrieved documents, N is the number of relevant documents, rel(k) is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

Finally, Mean Average Precision (MAP) for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{n} AP(q) \qquad (9)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

For both methods, we find average values of p(k) (with k=5,10,15) and compare the MAP values to examine the performance of proposed method with the mono-lingual information retrieval system. This is done because the performance of monolingual retrieval system is considered as an unreachable upper-bound of CLIR as translation process introduces translation error.

The MAP of our proposed method is 0.532 which is 85% of the baseline method. Thus our method achieves a comparable effectiveness with monolingual translation and is also much high than the 60% barrier limit of dictionary based query translation.

Scarcity of resources in Indian languages makes it quiet difficult to achieve efficient CLIR for these languages. Various authors have used different techniques for translation and comparison with them reveals the effectiveness of our method. Table 2 mentions language pairs, techniques used and their success rate (it is either mentioned in terms of Mean Average precision or percentage of monolingual retrieval). Comparative results show that our algorithm outperforms most of these methods.

Table 2: Success rate of Translation Technique used for Indian language pair

Language pair	**Translation Technique**	**Success Rate**
English to Hindi S. Varshney and J. Bajpai(2013) [22]	Bilingual dictionary	MAP is 0.3609
English to Hindi A.Seetha , S.Das & M. Kumar (2007) [23]	Select first equivalent/ preferred –n/ random nth equivalent/ all equivalents from Bilingual dictionary	MAP are 64.80%, 57.90%, 11.83% and 57.13% of monolingual retrieval
Tamil to English S. Saraswathi & A. Siddhiqaa(2010) [24]	Machine translation and Ontological tree	relevance improves only by 40% for English and 60% for Tamil
English to Hindi A.Seetha , S.Das & M. Kumar (2010) [25]	Bilingual dictionary and post query expansion	MAP is 0.0299
Hindi to English & Marathi to English M. Chinnakotla_, S. Ranadive, Om P. Damani, and P. Bhattacharyya (2008) [26]	Bilingual dictionary	For Hindi MAP of 0.2952 using title and description and for Marathi, we MAP of 0.2163 using title is achieved.
Hindi to English R. Udupa & J. Jagarlamudi (2008) [27]	Probabilistic translation lexicon produced by Statistical Machine Learning	Retrieval performance is about 81% of that of monolingual system
English to Hindi & Hindi to English S. Sethuramalingam & V. Varma (2008) [28]	Bilingual Dictionary	English-Hindi CLIR performance is 58% while Hindi-English CLIR is 25% of the monolingual performance
English to Bangla A.Imam & S. Chowdhury (2011) [29]	SMT using parallel corpus	NIST & BLUE scores (scoring system for evaluating the performance of a Machine Translation System.)are 4.6 and 0.39 which is below the standard

A. Observation:

Most of the cross lingual researches in Indian languages have used bilingual dictionary for query translation. But these lookups are independent of the context in which the term lies. A.Seetha et al. [23] have used three strategies

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

to obtain required translation form dictionary. They either select first equivalent/ preferred –n/ random nth equivalent/ all equivalents from Bilingual dictionary without considering the context of query term. M. Chinnakotla [26], though make use of mutual information between query terms they consider all terms equally important in the query thereby having less MAP as compared to our proposed system. A.Imam & S. Chowdhury [29] use parallel corpus to find translation. The results obtained are not encouraging as reported by them. This can be due to scarcity of parallel corpora for Indian languages. Our proposed method overcomes this problem by utilizing monolingual corpus which is still easier to build as compared to parallel corpus for Indian languages.
There can basically be three factors that make our algorithm better as compared to others. Firstly to disambiguate polysemous words, the algorithm relies on the context in which the term occurs, secondly it gives more weightage to discriminating terms in user query and thirdly it uses only monolingual corpora which is still easier to built as compared to parallel corpus for Indian languages.

V. Conclusion and Future Work

In this paper, we described our approach of query translation, which utilizes the concept of usefulness of context terms in finding the correct translation of target term. Our introduction of Weighted Term Similarity formula helps us in achieving comparable effectiveness with monolingual translation. As per the result comparison our method performs better than many other methods used for Indian languages CLIR.

In future we aim to test if the length of the query play any role in improving the mean average precision (MAP) of our proposed algorithm.

REFERENCES

1. P. Iswarya and V. Radha, "Cross Language Text Retrieval: A Review" .International Journal Of Engineering Research And Applications. 2(5), pp.1036-1043, 2012.
2. B. Ho, V.B. Dang, M.V. Luong, and T.T.B. Dong, "English-Vietnamese Cross-Language Information Retrieval: An experimental study". IEEE International Conference on Research, Innovation and Vision for the Future, pp 107-113, 2008.
3. R. Sperer and D. Oard, "Structured query translation for cross-language information retrieval" in Proceedings of the ACM SIGIR Conference. ACM, New York, 2000.
4. O.W. Kwon, et al., "Cross-Language Text Retrieval Based on Document Translation Using Japanese-to-Korean MT system" in Proc. of NLPRS'97, pp. 101-106, 1997.
5. D.A. Hull and G. Grefenstette, "Querying across languages: a dictionary-based approach to multilingual information retrieval" in Proc. of the 19th ACM SIGIR Conference (SIGIR'96), 1996.
6. Davis, M. "New experiments in cross-language text retrieval at NMSU's computing research lab" in Proc. of the fifth Text Retrieval Conference (TREC-5), 1996.
7. D. Eichmann, M.E. Ruiz, and P. Srinivasan, "Cross-Language Information Retrieval with the UMLS Metathesaurus". in Proc. of the 21th ACM SIGIR Conference (SIGIR'98), 1996.
8. M.G.,Jang, et al., "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting" in Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
9. J.H. Chun, "Resolving Ambiguity and English Query Supplement using Parallel Corpora on Korean English CLIR system". MS thesis, Dept. of Computer Science, KAIST (in Korean), 2000.
10. M. Adriani, "Using statistical term similarity for sense disambiguation in cross-language information Retrieval". Inf. Retr., 2(1):71–82, 2000.
11. K.W. Church, and P. Hanks, "Word Association Norms Mutual Information and Lexicography". Computational Linguistics, 16(1), pp.23-29, 1990.
12. L. T. Giang, V. T. Hung and H. C. Phap, "Experiments with Query Translation And Re-ranking Methods In Vietnamese-English Bilingual Information Retrieval". SOICT'13, Danang, Vietnam, December 05 – 06, 2013.
13. A. Maeda, F. Sadat, M. Yoshikawa and S. Uemura, "Query term disambiguation for web cross-language information retrieval using a search engine" in IRAL '00: Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages, pages 25–32. ACM Press, 2000.
14. P. Bajpai and P. Verma, "Query Term Disambiguation Using Co-occurrence Statistics for Dictionary based Cross Lingual Information Retrieval" in International Journal of Advances in Management, Technology & Engineering Sciences, Vol. IV, Issue 6(II), March 2015.
15. J. Gao, J. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. "Improving query translation for cross-language information retrieval using statistical models". Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp 96-104, 2001.
16. L. Ballesteros and W. B. Croft, "Resolving ambiguity for cross-language retrieval" in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 64-71, 1998.
17. J. Gao, et al., "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependency relations" in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 183–190, 2002.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

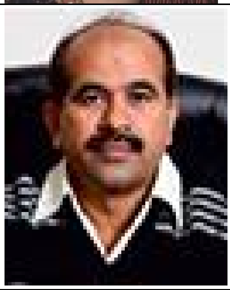
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

18. Y. Liu, R. Jin and J. Y. Chai, “A maximum coherence model for dictionary-based cross language information retrieval” in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’05). Salvador, Brazil. ACM Press, 536–543. 1076125, 2005.
19. D. Zhou, et al., “Disambiguation and Unknown Term Translation in Cross Language Information Retrieval”. Springer-Verlag Berlin Heidelberg (CLEF 2007), pp. 64–71, 2008.
20. A. Duque, L. Araujo, and R. Juan, “CO-graph: A new graph-based technique for cross-lingual word sense disambiguation” in Natural Language Engineering, Volume 21 , Special Issue 05 , pp 743-772, November 2015.
21. Salton G and McGill M.J. “Introduction to Modern Information Retrieval”. McGraw-Hill, New York, 1983.
22. S. Varshney and J. Bajpai, “Improving performance of English-Hindi cross language information retrieval using transliteration of query terms” in International Journal on Natural Language Computing (IJNLC) Vol. 2, No.6, December 2013.
23. A. Seetha, S. Das, and M. Kumar, “Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method”. In: Proceedings of 10th International Conference on Information Technology (ICIT2007).
24. S. Saraswathi, et al., “BiLingual Information Retrieval System for English and Tamil”, Journal Of Computing, 2,4, 85-89, April 2010.
25. A. Seetha, S. Das, and M. Kumar, “Post Translation Query Expansion using Hindi Word-Net for English-Hindi CLIR System”. In FIRE 2010, Gandhinagar, Gujrat, India, Month 2, 2010.
26. M.K. Chinnakotla, et al., “Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007”, in the working notes of CLEF 2007.
27. R. Udupa & J. Jagarlamudi, “Microsoft Research India at FIRE2008: Hindi-English Cross-Language Information Retrieval”. Working notes for Forum for Information Retrieval Evaluation (FIRE) Workshop, India 2008.
28. S. Sethuramalingam & V. Varma, “IIIT Hyderabad’s CLIR experiments for FIRE-2008”. The working notes of First Workshop of Forum for Information Retrieval, India 2008.
29. A.H. Imam, et al., “Impact of corpus size and quality on English-Bangla statistical Machine Translation system” in Proceedings of 14th International Conference on Computer and Information Technology (ICCIT), pages 566 – 571, Bangladesh 2011.

BIOGRAPHY

	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.