A Comparative Study of Techniques for Hindi-to-English Cross-Language Information Retrieval

Samra Kasim
Department of Computer Science, Whiting School of Engineering, Johns Hopkins University

***Abstract* – Culture and knowledge move seamlessly across digital borders, making it essential that Information Retrieval systems do so as well. This study evaluates the effectiveness of machine translation, multilingual embeddings, OOV transliteration, and pre- and post-query expansion in Hindi-to-English CLIR utilizing term, 4-gram, and 5-gram indices. 4-gram indices demonstrate the best CLIR performance among indices, especially for longer queries. The technical advances in machine translation tools result in better retrieval performance over word-to-word embedding translations. Thus, emphasizing the need for phrasal translation for improved CLIR performance. In addition, OOV transliteration and pre- and post-query translation showed effectiveness in enhancing embedding retrieval.**

I. Introduction

It is estimated that at least half the world's population is bilingual[1]. Today, it is not uncommon for a bilingual person to seamlessly cross language borders multiple times a day. A Netflix viewer can browse Hollywood and Bollywood movies at the same time. A k-pop fan can read reviews of their favorite albums in French or Korean. The task of retrieving documents in a language different from the query language is called Cross-Lingual Information Retrieval (CLIR). In our present day, when borders are no longer barriers to knowledge sharing, information retrieval (IR) systems must prioritize cross-lingual IR much the same as monolingual IR.

This study aims to leverage term, 4-gram, and 5-gram inverted indices to evaluate Hindi-to-English CLIR performance utilizing machine translation, multilingual embeddings, pre- and post-translation query expansion, and out-of-vocabulary (OOV) transliteration for query translation.

II. Related Work

The history of CLIR research can be traced back to 1970 and the work of Gerard Salton. Much has been written on this topic in the intervening half century. [1] Some of the more recent research of note is summarized in this section.

As outlined by Kishida and Chen, CLIR techniques typically consist of "two main modules: (1) [query] translation and (2) monolingual IR." [1]

In 1998, Ballesteros and Croft identified the "reduction of ambiguity" as one of the main challenges with query translation. Later, in 2002, McNamee and Mayfield demonstrated that combining pre- and post-translation query expansion improved precision and recall, but pre-translation had a larger effect. [2] Additionally, in 2004, McNamee and Mayfield

[1] https://www.washingtonpost.com/local/education/half-the-world-is-bilingual-whats-our-problem/2019/04/24/1c2b0cc2-6625-11e9-a1b6-b29b90efa879_story.html

demonstrated that "overlapping character n-gram tokenization can provide retrieval accuracy that rivals the best current language-specific approaches for European languages", and concluded that 4-gram and 5-grams performed well for all languages. [3]

In 2001, Chen demonstrated that pre-query phrasal translation outperformed word translation by 23%, while in 2002, Diekema concluded that there are two main causes of query translation error: (1) ambiguity and (2) "word-by-word translation of terms that are part of multiple-word expressions." [4]

Due to technical advances, machine translation systems are often employed for query translation. Vahid et al examined the effectiveness of Google and Bing translation for CLIR, concluding that "CLIR performance is affected by machine translation effectiveness in different language pairs but also in the same language pair for different query sets." [5]

In 2016, working with Hindi-to-English CLIR, Bhattacharya et al, demonstrated a word embedding based approach that outperformed basic dictionary translation by 70% and a hybrid dictionary-embedding approach that beat the dictionary baseline by 77%. The embeddings, when combined with Google Translate and dictionary, bested the monolingual baseline by 15%. [6]

III. Materials and Method

The following section discusses the materials and methods undertaken to evaluate CLIR performance. The English language 2020 NIST TREC COVID shared task corpus is leveraged for this study.

a. Multilingual Embedding Generation

Leveraging English and Hindi pre-trained aligned word vectors in FastText[2] format, Hindi-English and English-Hindi cross-lingual word embeddings are generated with Vecmap[3] , an open-source tool designed for this purpose. Additionally, Facebook's MUSE[4] bilingual Hindi-to-English dictionary is employed for semi-supervised training on GPU with Vecmap.

b. Inverted Index Creation

For this study, term, 4-gram, and 5-gram indices are generated. For each document, stop words and punctuation are removed utilizing NLTK. For the term index, documents are tokenized into words utilizing NLTK's word_tokenize method. Digits are included in the index to enable results for queries such as "1600 Pennsylvania Ave" where digits are essential. For the 4-gram and 5-gram indices, each word token is converted into its resultant grams. For each term, a postings list is generated. The postings list is a list of documents in which the term occurs and the frequency of that occurrence (i.e., term

[2] https://fasttext.cc/docs/en/aligned-vectors.html
[3] https://github.com/artetxem/vecmap
[4] https://github.com/facebookresearch/MUSE

frequency). The postings are written to a binary file to minimize disk space and memory usage. Additionally, each term is stored in a Python dictionary as key with an offset pointer to the binary file as value. Thus, only the dictionary is loaded into memory for querying. The term's offset is utilized to perform read operations on the binary file to load the postings list for a term into memory only when it is needed.

c. Query Translation

The 2020 NISST TREC COVID shared task consists of queries in two formats: (1) keywords and (2) questions. Since the original queries are in English, to generate the baseline Hindi queries, the English queries are translated from English to Hindi utilizing Google Translate and then manually corrected for deficiencies and errors.

Utilizing the Hindi queries, the following translation mechanisms are employed to generate English queries from the Hindi queries:

1. **Machine Translation:** Hindi language queries are translated to English using Google Translate
2. **Cross-Lingual Embeddings**: Cross-Lingual Embeddings generated with Vecmap are employed to translate Hindi words to English
3. **Cross-Lingual Embeddings with Out-of-Vocabulary (OOV) transliteration:** Hindi words such as कोरोनावाइरस (coronavirus) are OOV and were not translated with the Cross-Lingual embeddings. Consequently, these OOV words were left out of the previously mentioned query. To mitigate this, a transliteration tool was developed for this research to transliterate Hindi words to Romanized text. As such, the above is transliterated to koronavaars.
4. **Cross-Lingual Embeddings with OOV transliteration and pre- and post-translation query expansion:** Queries generated using the pervious method are further expanded. For pre-query translation, FastText's monolingual Hindi embeddings are utilized to return the top 3 most similar words to each query word and are added to the query. Additionally, for post-query translation, the cross-lingual Hindi-to-English trained embeddings are utilized to return the top 3 most similar words to the translated query word, which are then added to the translated query.

d. Ranked List Generation

To generate the ranked list, each query is parsed and tokenized in the same manner outlined above (part b) for documents. Terms/n-grams are weighted using term frequency-inverse document frequency (TFxIDF) score, which is a statistical measure that evaluates how important a term/n-gram is to a document in a collection[5]. As a result, terms/n-grams occurring in many documents are weighted less, while terms that occur rarely in documents are weighted more.

The IDF score is calculated as follows:

[5] https://en.wikipedia.org/wiki/Tf-idf

$$IDF(t) = log_2 \frac{df(t)}{N}$$

The TFxIDF score is employed to get real-valued vectors[6] for queries and documents. Then, utilizing the cosine similarity equation below, the measure of similarity between the two vectors is returned.

$$cos\theta = \frac{d \cdot q}{|d||q|}$$

This similarity measure is employed to rank documents. Document vectors and query vectors that show the most similarity are returned higher in the ranked list. For this study, the top 100 scores for each query are returned. In addition to the queries listed in part c, the original English queries are also used to generate a ranked list, which serves as the monolingual baseline to measure other results.

e. Ranked List Evaluation

The ranked lists are evaluated with TREC_EVAL[7], which uses standard NIST evaluation procedures for IR systems[8] and generates key performance metrics such as Mean Average Precision (M.A.P), precision at k, and recall.

IV. Results and Analysis

Utilizing M.A.P scores provided by TREC_EVAL, the results of experiments conducted on term, 4-gram, and 5-gram indices are summarized in this section.

As expected, the monolingual IR experiment had the best performance across indices for keyword and question queries. Additionally, M.A.P scores were higher for the questions queries than for keyword queries. This is an expected result because longer queries have better results because more words in a query mean more potential matches for a document. This in turn generates a higher cosine similarity score between a document vector and query vector.

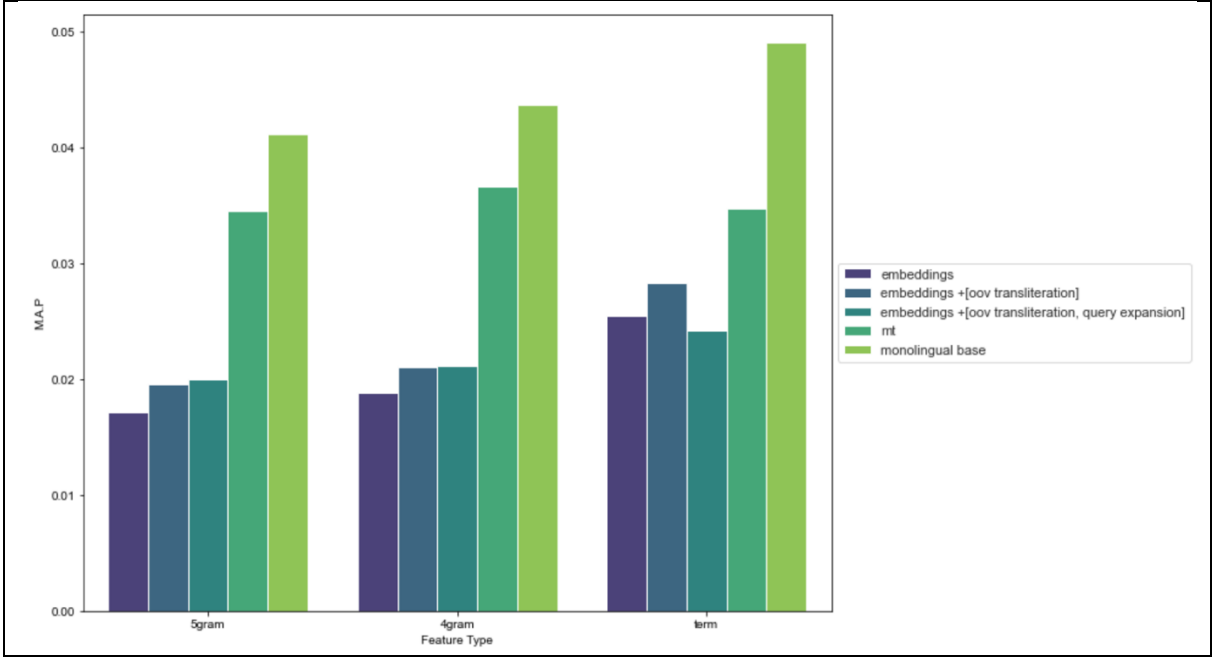
However, for keyword queries, the top M.A.P score was for the term index. Contrast this with queries where the top score for the monolingual baseline was in the 4-gram index. Examples of keyword queries are “serological tests for coronavirus” and “how do people die from the coronavirus”. Once stop words are removed, the keyword queries are short. On the other hand, examples of question queries include “are there serological tests that detect antibodies to coronavirus?” and “what drugs have been active against SARS-CoV or SARS-CoV-2 in animal studies?” N-grams create a normalized representation of text that handles morphological variation. It follows, then, that longer sentences in the question

[6] https://stackoverflow.com/a/6255993/4882806
[7] https://github.com/usnistgov/trec_eval
[8] http://faculty.washington.edu/levow/courses/ling573_SPR2011/hw/trec_eval_desc.htm

queries are more likely to have more morphologically complex text that benefits from n-gram representation than keyword queries.

Figure 1: Keywords Queries: M.A.P. as percentage of monolingual English-English Query			
Experiment	**Term**	**4-gram**	**5-gram**
Monolingual Baseline	.0491**	.0437	.0412
Machine Translation	.0348 (70.89%)	.0366** (83.75%)	.0345 (83.74%)
Cross-Lingual (CL) Embeddings	.0255** (51.93%)	.0188 (43.02%)	.0172 (41.75%)
CL Embeddings +[OOV transliteration]	.0283* ** (57.65%)	.0211 (48.28%)	.0196 (47.57%)
CL Embeddings +[OOV transliteration, pre- and post- translation query expansion]	.0242** (49.29%)	.0212* (48.51%)	.0200* (48.54%)
* top embedding experiment M.A.P score by index type (i.e., column) ** top M.A.P score by experiment (i.e., row)			

Figure 2: Keywords M.A.P.

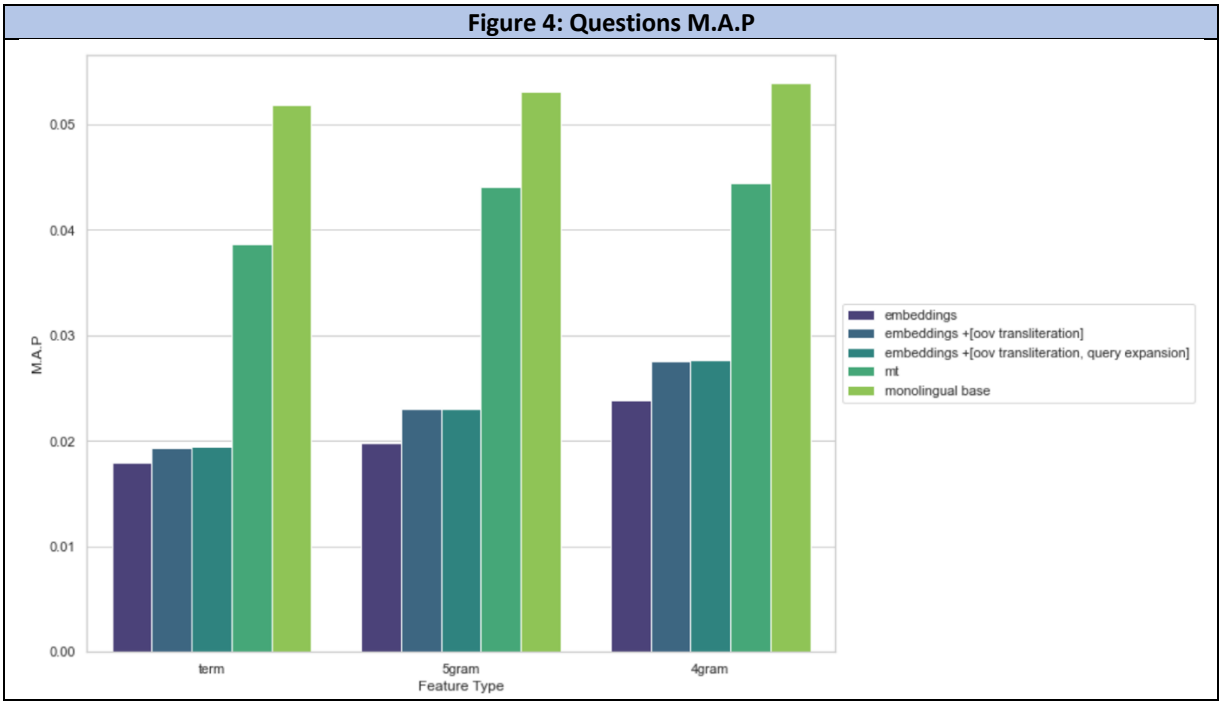


In the keyword queries and question queries, machine translation (using Google Translate) resulted in the second-best M.A.P scores across indices. Machine translation queries performed better than queries generated via cross-lingual embeddings. These results point to the technical advancements in machine translation tools. Further, embeddings utilized for this study leverage word piece encoding, while the machine translation tools are utilizing the full context of a sentence to generate the translation. Machine translation's advantage in phrasal translation explains the better performance over embeddings. For example, in the phrase "she will lead him to the market", it is more likely that a machine translation tool will parse lead as a verb and not as a chemical element.

For experiments with multilingual embeddings, the sole use of embeddings for translation demonstrated the worst M.A.P. scores. This was due to OOV words. Since the query sets had very specific COVID-19 and other scientific vocabulary, if a Hindi word was not found in the cross-lingual embeddings, then it was left out of the translated query.

Figure 3: Questions Queries: M.A.P. as percentage of monolingual English-English Query			
Experiment	**Term**	**4-gram**	**5-gram**
Monolingual Baseline	.0519	.0539**	.0531
Machine Translation	.0387 (74.57%)	.0445** (82.56%)	.0441 (83.05%)
Embeddings	.0180 (34.68%)	.0239** (44.34%)	.0198 (37.29%)
Embeddings +[OOV transliteration]	.0194 (37.38%)	.0276** (51.21%)	.0230* (43.31%)
Embeddings +[OOV transliteration, pre- and post- translation query expansion]	.0195* (37.57%)	.0277* ** (51.39%)	.0230* (43.31%)
* top embedding experiment M.A.P score by index type (i.e., column) ** top M.A.P score by experiment (i.e., row)			

Figure 4: Questions M.A.P

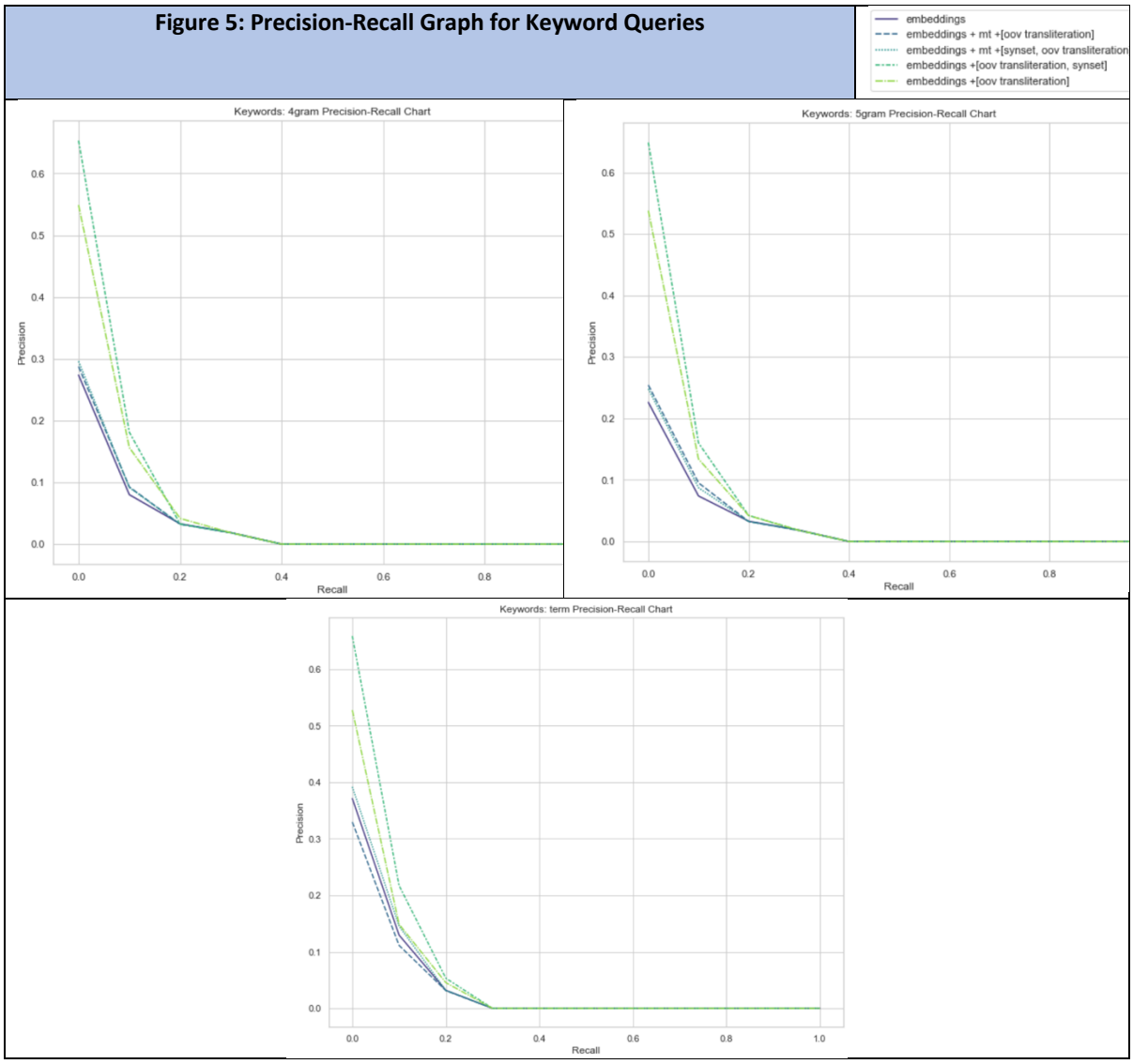


To improve on the previous query, OOV transliteration was added to the translated queries. The transliteration tool was developed specifically for this study and it transformed words in Devanagari script to romanized text. As a result, सीरोलॉजिकल (serological) was transliterated as sirolojikl. Proper nouns benefitted the most from transliteration. The n-gram queries also saw an improvement with the addition of transliteration likely because even if a transliterated word was not an exact match for its English translation, there was enough overlap that it was captured by n-grams between the translated word and the transliterated word.

For the keyword queries, embeddings with OOV transliteration demonstrated the best performance for the term index (including over embeddings with OOV translation and pre-

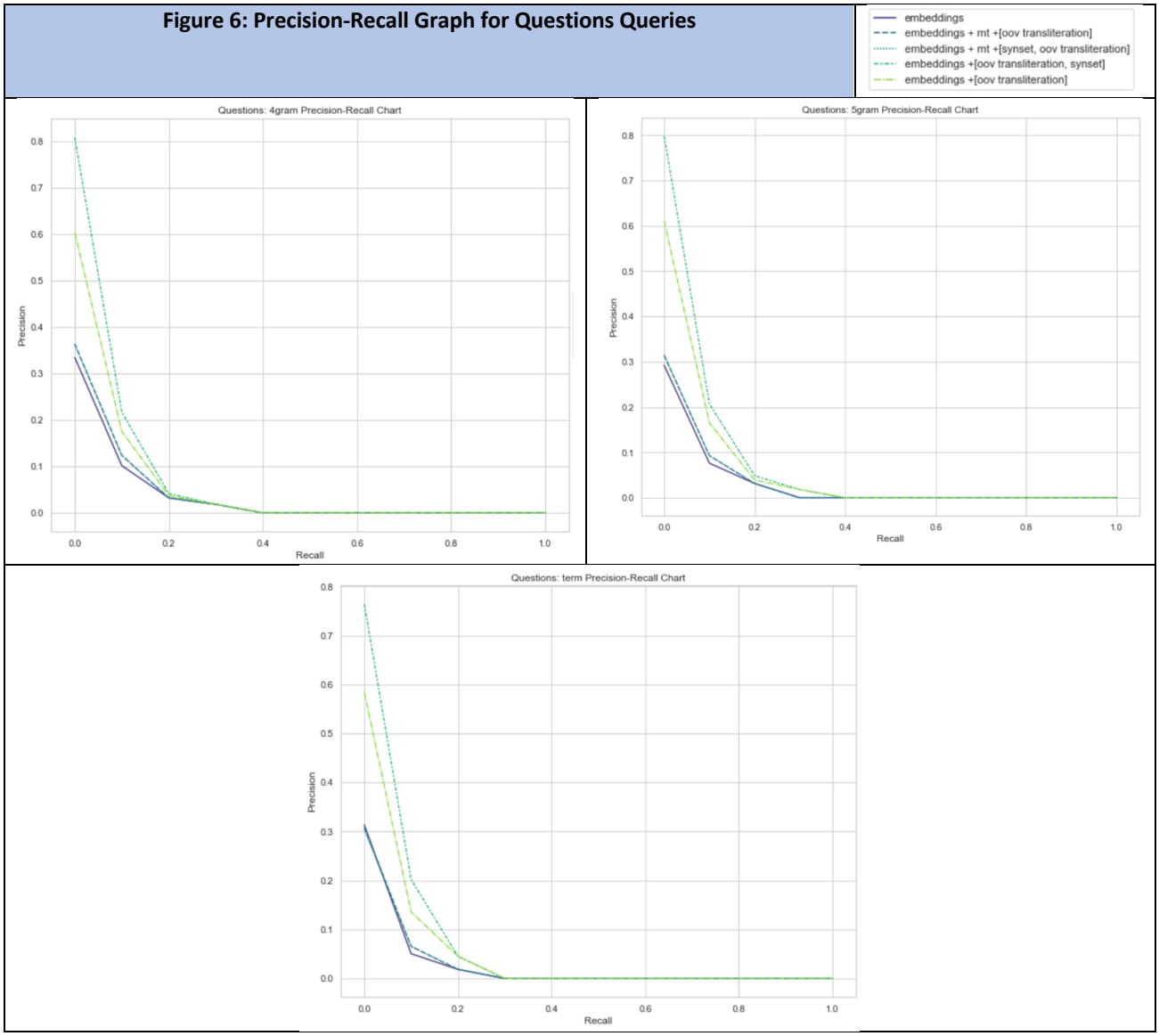
and post-translation query expansion), likely because query expansion diluted the intent of the original query.

Figure 5: Precision-Recall Graph for Keyword Queries



Embeddings with OOV translation and pre- and post-translation query expansion demonstrated the best performance for 4-gram and 5-gram indices for both query formats and for the term index for question queries. Overall, 4-grams demonstrated the top M.A.P scores in this category. For instance, the question query that in the original English reads “are there serological tests that detect antibodies to coronavirus?” was expanded with pre- and post-translation query expansion and transliteration to “what wondering how where wherever nearby sirolojikl testing tests test are does is which the it koronvayrs to them instead antibody antibodies antibody/antigen the of a confirm confirms reveal puts comes pushes hai”. As evidenced by the morphological variations for test and antibody, the expanded queries are more likely to capture the intent of the query rather than the exact words and thus match documents that otherwise may have been missed.

Figure 6: Precision-Recall Graph for Questions Queries



The eleven-point interpolated average precision-recall graphs for both query formats in Figures 5 and 6, confirm the findings outlined above using M.A.P. scores. Notably, the question queries have a greater area under the curve than the keyword queries representing both higher recall and higher precision than the latter query format.

V. Conclusions and Future Work

This study investigated the effectiveness of machine translation, embeddings, OOV transliteration, and pre- and post-translation query expansion in Hindi-to-English CLIR. Machine translation demonstrated the best results when compared to the monolingual baseline because the translation tool effectively captured phrasal context in a query. Without phrasal context, word-by-word translation leveraged by embeddings suffered significantly in performance likely due to sense ambiguity in query terms. The addition of OOV transliteration and pre- and post-

translation expansion to embedding translation improved performance. 4-gram index had the best performance for question queries due to the longer length of the queries as well as more morphological variety in the text.

To improve upon this study, future work should focus on phrasal translation and sense disambiguation. This study only utilized within word n-grams and the results may be improved by utilizing across word n-grams to capture phrasal context. Additionally, indices leveraging bigrams or trigrams may also address the issue of phrasal context in translation. Finally, statistical methods to identify noun phrases have proven successful in other studies and can be employed to improve query translation. [7]

References

[1] Kishida K., Chen K. (2021) Experiments on Cross-Language Information Retrieval Using Comparable Corpora of Chinese, Japanese, and Korean Languages. In: Sakai T., Oard D., Kando N. (eds) Evaluating Information Retrieval and Access Tasks. The Information Retrieval Series, vol 43. Springer, Singapore. https://doi.org/10.1007/978-981-15-5554-1_2

[2] McNamee, P., & Mayfield, J. (2002). Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 159–166). Association for Computing Machinery.

[3] McNamee, P., Mayfield, J. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* **7**, 73–97 (2004). https://doi.org/10.1023/B:INRT.0000009441.78971.be

[4] Diekema, A. (2002). Translation Events in Dutch Cross-Language Information Retrieval. *ITLS Faculty Publications*. https://www.researchgate.net/publication/44883670_Translation_Events_in_Dutch_Cross-Language_Information_Retrieval

[5] Hosseinzadeh Vahid, A., Arora, P., Liu, Q., & Jones, G. (2015). *A Comparative Study of Online Translation Services for Cross Language Information Retrieval*. https://www.researchgate.net/publication/280553348_A_Comparative_Study_of_Online_Translation_Services_for_Cross_Language_Information_Retrieval

[6] Paheli Bhattacharya, Pawan Goyal, & Sudeshna Sarkar. (2016). UsingWord Embeddings for Query Translation for Hindi to English Cross Language Information Retrieval. https://arxiv.org/abs/1608.01561

[7] Gao, J., Nie, J.y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. https://www.researchgate.net/publication/2405090_Improving_Query_Translation_for_Cross-Language_Information_Retrieval_using_Statistical_Models