# 605.744: Information Retrieval
# Problem Set (Module 8)

Sabbir Ahmed

October 20, 2022

1. (20%) Name three significant issues that arise when using dictionaries to translate queries in cross-language information retrieval and briefly explain why they create a problem.

   **Answer:** Some of the significant issues of using dictionaries in CLIR are: out-of-vocabulary words, processing conjugations and lexical ambiguity. Out-of-vocabulary words are words that do not exist in the dictionaries for the target language. The IR system therefore does not know how to translate these words. The dictionaries may have the base form of a word, but not its numerous conjugations, which will lead to the IR system failing to recognize the various forms. Lexical ambiguity refers to the possibility of a single word having numerous meanings depending on the context, and translating them directly may lead to errors in semantics.

2. (20%) Give two advantages and one disadvantage of using character n-gram tokenization for multilingual text retrieval.

   **Answer:** One disadvantage of character n-gram tokenization is the increased requirements in computation time and storage capacities. A token with $m$ characters get replaced with $m - n + 1$ tokens. The advantages of using character n-grams are that they can be language independent, since tokenization involves simple, fixed algorithms. Another benefit is that it can help address issues with conjugations of words, where several n-grams are shared between the different forms of the base word.

3. (20%) For this question consider an English alphabet to consist of just 26 (lower-cased) letters, 10 digits, and a space character. And consider there to be exactly 10,000 characters in Chinese. Note, spaces are not used in written Chinese.

   (a) How many possible character 4-grams are there in English? Using Table 5.1 (IIR) how does this number compare to a typical vocabulary size when words are used?

   **Answer:** Assuming the English alphabet consist of 26 lower-cased letters, 10 digits and 1 space character, that totals to 37 characters. The possible number of character 4-grams is $37^4 = 1,874,161$. This value is considerably larger than the vocabulary sizes listed in Table 5.1, regardless of the levels of filtering.

   (b) How many possible indexing terms will there be if 2-gram indexing is used for Chinese? What if 3-grams are used?

   **Answer:** Assuming the Chinese alphabet consist of 10,000 characters, the possible numbers of 2-grams and 3-grams are $10,000^2 = 1 \times 10^8$ and $10,000^3 = 1 \times 10^{12}$ respectively.

(c) What difficulties might occur when indexing a document collection if the vocabulary size (i.e., number of indexing terms) is extremely large?

**Answer:** Some common issues of indexing an extremely large vocabulary size is the time and space required to generate such an index file.

4. (20%) What advantages does query translation have over document translation in cross-language information retrieval (CLIR)?

   **Answer:** Although documents can improve the translation with its contexts, they are generally much larger than queries. The main advantage of translating queries is that they are lighter in terms of computation requirements.

5. (20%) Briefly describe what pre-translation query expansion (sometimes called pre-translation feedback) is and then explain why it is helpful in dictionary-based cross-language information retrieval.

   **Answer:** Query expansion during relevance feedback refers to suggesting users to input additional query terms. Expanding a query prior to translation and using it to search in the target collection can improve its performance. This method can be useful in tackling the major issues in dictionary-based CLIR, such as disambiguating the query and resolving out-of-vocabulary words.