Sheetal Parikh
Problem Set – Module 8
605.744.81 Information Retrieval
Fall 2021

1.) 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.

In cross-language information retrieval the language of the query is different than the language of the document collection. When using dictionaries to translate the queries, the query terms are mapped into the target language of the document collection. Dictionaries can contain a list of words and their translations. One issue that can arise is untranslatable search keys, meaning that you have out of vocabulary words or you have words missing from the translation dictionary. If words are not included in the dictionary, then it cannot be translated causing incorrect translations. Another issue that can arise is the processing of inflected words, which occurs when the dictionary has only the base form of a word but does not have conjugated forms of the word (for example, if the dictionary contained the word build but not built or building). Similarly to untranslatable search keys, certain words that are a different version of the word included in the dictionary, will not be translated causing incorrect translations. A third issue that may arise is phrase identification and translation which occurs when the dictionary does not have translations for multiword phrases. Each word may be translated individually but the entire phrase cannot be translated and so the meaning of the phrase may be lost, causing incorrect translations.

References:
Lecture 8D
Pages 324 – 325: https://link.springer.com/content/pdf/10.1023/A:1011994105352.pdf
Page 325:
https://www.researchgate.net/publication/282245668_Dictionary_Based_Translation_Approaches_in_Cross_Language_Information_Retrieval_State_of_the_Art

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

Disadvantage

A disadvantage of using character n-gram tokenization for multilingual text retrieval is that if n-gram tokenization is done naively it can take up more space in an inverted file and dictionary and thus have negative speed and disk space costs compared to just using words as indexing terms.

Advantages

An advantage of character n-grams is that it is an overall simple technique and can be applied to any language. Another advantage is that single spelling errors may only effect some of the n-gram terms. Therefore, you will not completely lose a word if you have a spelling error.

References:
Lecture 8B

3.) 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?

Assuming that letters, space, and characters are all tokens we have a total of 37 characters in the text:

26 letters + 10 digits + 1 character space = 37

Text: abcdefghijklmnopqrstuvwxyz0123456789_

We will use the text above to create the following 4-grams:

1. abcd
2. bcde
3. cdef
4. defg
5. efgh
6. fghi
7. ghij
8. hijk
9. ijkl
10. jklm
11. klmn
12. lmno
13. mnop
14. nopq
15. opqr
16. pqrs
17. qrst
18. rstu
19. stuv
20. tuvw
21. uvwx
22. vwxy
23. wxyz
24. xyz0
25. yz01
26. z012
27. 0123
28. 1234
29. 2345
30. 3456
31. 4567
32. 5678
33. 6789
34. 789_

We get a total of 34 4-grams.

The formula for the total number of n-grams can be used as well: m-n+1, where
m = the length of the text and n = the number of the fixed n-gram.

m = 37
n = 4
total number of 4-grams = 37 - 4 + 1 = 34

Table 5.1, shows the effect of stop word removal and stemming on both the dictionary and index size using Reuters-RCV1. When looking at the unfiltered row, the text has 484,494 terms, 109,971,179 non-positional postings, and 197,879,290 tokens. When looking at the typical vocabulary size when words are used, as in the table, and showing how when using a small text of only 37 in length, we can see that if a very large corpus would be used with character n-grams, we would generate a much larger number of indexing terms than what's shown in the table. The table helps bring into perspective the difference in indexing terms that would have to be processed when using words vs. character n-grams.

References:
Lecture 8B

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

Length of text for Chinese = 10,000 (no spaces)

Using the formula above, m – n + 1, we get the following number of indexing terms if using 2-grams:

m = 10,000
n = 2
Total number of 2-grams = 10,000 – 2 + 1 = 9,999

Using the formula above, m – n + 1, we get the following number of indexing terms if using 3-grams:

m = 10,000
n = 3
Total number of 3-grams = 10,000 – 3+ 1 = 9,998

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

When indexing a collection using character n-grams, if a word or phrase occurs frequently in a collection, it will also have many corresponding n-grams. Also, words with the same root may generate many of the same n-grams. Therefore, if the vocabulary size is extremely large, indexing a document collection using n-grams will create significant redundancies that would require an increased disk space for all the indexed terms. This would also result in and high run-time costs.

References:
Page 1: https://arxiv.org/pdf/1806.09447.pdf
Page 5: https://users.soe.ucsc.edu/~elm/Papers/jodi00.pdf

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

There are two primary approaches in CLIR, query translation and document translation. In document translation we are translating documents into the query language, whereas for query translation we are translating the query into the document language. If we do not know the query language ahead of time, document translation can be very costly and time consuming as it would require translating the documents into all possible query languages and then indexed. This would also make document translation very hard to scale. However, translating only the query would save time and be less computationally expensive as only the query needs to be translated which should be a much shorter text than the documents. Since query translation is a faster process, it also is more flexible and can allow more interaction with the users, if the user understands the translation.

References:
Lecture 8D
Page 6849: https://aclanthology.org/2020.acl-main.613.pdf

Pages 208 – 209: https://aclanthology.org/P99-1027.pdf
https://medium.com/lily-lab/a-brief-introduction-to-cross-lingual-information-retrieval-eba767fa9af6

5.) 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.

Pre-translation query expansion is another method for improving bilingual retrieval. The original query is expanded using a set of feedback documents retrieved in the source language. The expanded query is then translated and then used for searching in the target collection. This process uses the entire query and adds terms related to the query improving the probability of the matching term and limiting the possibility of failing to translate a term or phrase in the query. Because the entire query is used and pre-translation feedback provides a wider range of expressions, pre-translation query expansion can help translation ambiguity or when the word can be translated into multiple possible variants. Subsequently, this may help improve precision.

References:
Lecture 8D
Pages 1-2, 5: https://faculty.washington.edu/levow/papers/IRAL03.pdf