Cross-language Retrieval In English and Vietnamese

Van Be Hai Nguyen and Ross Wilkinson and Justin Zobel
Department of Computer Science, RMIT
Melbourne, Australia
{behai,ross,jz}@mds.rmit.edu.au

Abstract
A cross-lingual Vietnamese-English retrieval system should enable Vietnamese speakers to retrieve documents written in English, and likewise English speakers should be able to retrieve documents written in Vietnamese. We have conducted preliminary experiments to investigate such possibilities. Using a simple language transformation algorithm that we have developed, our results indicate that cross-language retrieval for English and Vietnamese using bilingual dictionaries is feasible.

Introduction
Standard information retrieval systems are monolingual, providing access to documents in a given language via queries expressed in that language. In cross-lingual retrieval, the queries and documents are in different languages, so that some form of translation or transformation is required for queries to be resolved.

We explore two issues through the experiments reported in this paper: how different levels of automatic transformation affect retrieval performance, and the limits of effectiveness of cross-lingual retrieval.

Experiments
The test collection used in the experiments is the standard TIME [5] collection, which has 83 queries and associated relevance judgements. This collection is about world affairs in the 1960s and contains 423 documents (totalling about 1.6 Mb of text). We have had all the documents and queries in this collection manually translated into Vietnamese, by native speakers who however are not professional translators.

The transformation method we chose to explore was substitution of words in one language for words in another based on a bilingual dictionary. The bilingual dictionaries used in these experiments (one for English to Vietnamese and one for Vietnamese to English) are those compiled by Nguyen [3]. In their raw format, they are not suitable for such language transformation, so we have filtered them, eliminating all irrelevant information. This format we call original. By using the Lovins stemming algorithm [2] (for English) to stem the entries in the English-Vietnamese dictionary, the entries with the same root can be merged; this version we call Lovins. The Vietnamese-English dictionary has 36,187 entries; the original English-Vietnamese dictionary has 25,807 entries; and the Lovins English-Vietnamese dictionary has 21,426 entries.

The language transformation algorithm is simple. Each word in the source text is read, located in the bilingual dictionary, then replaced by the corresponding term, or terms. Vietnamese shares some characteristics with Chinese: segmentation is required to determine which phoneme sub-strings are words. To transform Vietnamese to English, probable words in the Vietnamese text are identified using a greedy segmentation algorithm described by Nguyen and Wilkinson [4].

For transformation of English and Vietnamese, words can be located in the original or Lovins dictionary. As another option we have also developed a stemming algorithm to progressively stem English affixes on the fly, based on the algorithm given by Kelly and Stone [1]. This stemmer reduces morphological words to their supposedly correct roots; for example, applied is changed to apply. This dictionary stemmer is used with the original English-Vietnamese dictionary.

In summary, there are three methods for transforming English to Vietnamese. Each of these methods produces different automatic transformations in Vietnamese. We show below how these differences affect retrieval performance. The hypothesis is that the Lovins and dictionary stemmer transformations will produce better retrieval results than transformation based on the original dictionary.

Experiments reported in this paper do not include automatically transforming Vietnamese queries or documents to English. Because the TIME documents and queries were manually translated to Vietnamese by non-experts, we do not believe that automatically
Results

The experiments investigate two issues: how the different automatic transformations affect retrieval performance; and the effectiveness of cross-language retrieval.

For English retrieval, we compared 12 variations, for example: stop, stop and stem, adjacent pairs and so on. For Vietnamese, we used both word-based indexing and morpheme-based indexing. We also tried several different variations as for English. We tried using different stop lists, using adjacent pairs, case folding, and different segmentation for the word-based runs. This led to 30 runs for both word-based and morpheme-based retrieval. For transformation, we conduct experiments using only the first corresponding term, and with all the corresponding terms in the bilingual dictionary. All results reported in this paper are the best of these runs for each of the experiments.

Retrieval performance with each of the transformation methods is reported in Table 1. As a baseline we have also tested monolingual retrieval: English queries on English documents gives retrieval effectiveness of 49.4% and Vietnamese queries on Vietnamese documents gives 41.3%. These numbers are 11-point average recall-precision. There is another important baseline: queries in English against Vietnamese documents. Note that since proper names are directly substituted when manually translating the documents, and given the high significance of proper names, we get a significant retrieval performance. In fact the result is an average retrieval effectiveness of 33.7%. This is a very high minimum! To investigate how effective transformation is without the masking effect of proper nouns, we conduct four further experiments with all proper removed from queries, with results reported in Table 2.

The Lovins method produces the worst results, probably because irrelevant words are combined in the Lovins dictionary, so it acts like a thesaurus that has unrelated words grouped together.

The original and the dictionary stemmer transformations are effectively indistinguishable. A possible explanation is that our dictionary stemmer is not effective; it recognises only a small number of affixes (the 16 prefixes anti-, co-, dis-, ex-, multi-, non-, it-, inter-, in-, ir-, mid-, pan-, pre-, re-, and un-; and the 18 suffixes -ism, -ive, -able, -ion, ization, -ment, -ity, -ably, -ness, -ic, or, -ed, -ics, -s, -er, -est, -ing, and -ly).

Monolingual English retrieval (49.4) is better than Vietnamese retrieval (41.3). Examination of the English queries that perform better than their Vietnamese counterparts shows that keywords are used more consistently in English documents and queries than in Vietnamese. For example, consider query 53:

PREMIER

KHRUSHCHEV ONCE AGAIN PRESSING FOR A NONAGGRESSION PACT BETWEEN NATO AND THE WARSAW PACT NATIONS

"Pact" is not present in the Vietnamese version of the relevant document because in the query the expression "hiếp định" is used, but in the document the concept is expressed as "hiếp uốc"; neither is incorrect—the usage is analogous to "treaty", "pact", and "agreement" in English—but the difference means that document and query do not match. "Aggression" also has two Vietnamese equivalents, "xâm lăng", which is seen in the query, and "xâm phạm", which is used in a relevant document.

The Time collection contains many proper nouns in both queries and documents. We extracted 2558 unique non-translatable proper nouns (mostly personal names, names of places, and so on) from the collection, and added them to the English-Vietnamese dictionaries prior to transformation from English to Vietnamese. In effect, all the proper nouns that appear in English queries and documents also appear in the Vietnamese transformations of these queries and documents. Thus variation in performance is due to other factors.

In the case where only queries or documents are automatically transformed, the performance is a little bit lower than when both are automatically transformed. This is probably because, apart from the proper nouns, they have different lexicons.

1 Conclusion

Our results show that cross-language retrieval is feasible for English-Vietnamese, indicating that, with a simple automatic transformation process applied to queries, we still get acceptable performance, with only a limited loss as compared to retrieval in the original languages. However, the preponderance of proper nouns in the queries is the likely reason that cross-lingual retrieval has worked so well.

Our next goal is to understand what lies behind these results. Our long term goal is to combine cross-language retrieval and machine translation to provide an information retrieval system that is capable of effectively handle cross-language retrieval.

References


<table>
<thead>
<tr>
<th>Transformation method</th>
<th>Original</th>
<th>Dict. Stem.</th>
<th>Lovins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese query vs English document transformed to Vietnamese</td>
<td>38.4</td>
<td>37.9</td>
<td>20.7</td>
</tr>
<tr>
<td>English query transformed to Vietnamese vs Vietnamese document</td>
<td>34.3</td>
<td>36.5</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Table 1: Cross-lingual retrieval performance with the standard TIME collection.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16.0</td>
<td>13.6</td>
<td>11.3</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Table 2: Retrieval effectiveness with no proper nouns.

