Integrating Part-of-Speech Constraints in Cross-Lingual Information Retrieval

Neha Verma and Dragomir Radev
Yale University
LILY Lab

Introduction

Cross-lingual information retrieval (CLIR) is the task of returning documents relevant to a given query where the language of the document and query differ. In this project, we are specifically interested in the task of CLIR involving low-resource languages. Within a set of queries, one might occasionally encounter ambiguous query terms with multiple senses. These terms can be disambiguated with their appropriate part-of-speech. Therefore, we seek to integrate query-level part-of-speech information into our retrieval system, in order to return documents relevant to the sense of the query term provided. Specifically, in the query language, we apply the part-of-speech constraint, and search for documents with this translated word with the correct part-of-speech. We additionally attempt translating the query terms, and applying the part-of-speech constraint in the target language, although this is not the exact specification of our query set.

Materials and Methods

All of the data used in this project was provided by IARPA for the MATERIAL (MAchine Translation for English Retrieval of Information in Any Language) program. We consider the retrieval setting between English queries with Farsi documents. We use Query pack 1, and the DEV document collections for this analysis.
We are provided with gold-label relevance annotations between the DEV set and Query pack 1, and use this in our analysis. Additionally, we are provided with English translations of the Farsi documents, from collaborators at other universities. Finally, we also have part-of-speech tags for each of these translated documents, provided also by collaborators. Since the translations and tags are not ground truth data, our work is subject to any errors from our translation models and part-of-speech taggers.
We use the Actual Query Weighted Value (AQWV) as our evaluation metric, as defined by IARPA. This metric penalizes both missed documents (false negatives) and, to a much smaller extent, false alarms (false positives).

$$QWV = 1 - \left(\frac{\sum_{j=1}^{NQ_{rel}} P_{Miss}(Q_j)}{NQ_{rel}} + \beta \frac{\sum_{j=1}^{NQ} P_{FA}(Q_j)}{NQ} \right)$$

$$P_{Miss} = \frac{N_{Miss}}{N_{rel}} \qquad P_{FA} = \frac{N_{FA}}{N_{total} - N_{rel}}$$

Figure 1. QWV computation, from IARPA. Beta is set to 40 in this evaluation period.

Collection	Count
Total queries	221
POS-constrained queries	89
Noun query terms	91
Verb query terms	2
Adjective query terms	13

Table 1: Query Pack 1 parts-of-speech distribution

Best System	MT Basic	MT Query Expansion	PSQ Indri
Original QWV (dev)	0.228	0.221	0.122
Filtered QWV (dev)	0.167	0.162	0.074

Table 2. QWV values for filtering documents with no POS match in target language (Farsi)

Best System	MT Basic	MT Query Expansion	PSQ Indri
Original QWV (dev)	0.228	0.221	0.122
Filtered QWV (dev)	0.167	0.162	0.065

Table 3. QWV values for filtering documents with no POS match in source language (English)

Results

In one experiment, we filter out all documents where the translated query term does not appear in the part-of-speech tagged Farsi documents. These results appear in Table 2. In the next, we filter out all documents where the POS-constrained query-term does not appear in the English-translated POS-tagged documents. These results appear in Table 3. As seen in both tables, filtering out documents does not improve the QWV score. In a fine grained analysis, filtering out such documents does improve the probability of a false alarm, as seen in Figure 1. However, we deduct from QWV due to filtering out documents that are actually relevant. In examining the denominators of the probability of a missed document and the probability of a false alarm in Figure 1, we see that the denominator of P_{miss} is much smaller than the denominator of P_{FA}. Therefore, we seek an approach that removes less relevant documents, and focuses on adding relevant documents.

Future Work

Because our baseline filtering approach proved to be harsh in terms of false negatives, we focus our future attention on 1) adding documents to the retrieval list based on POS information and 2) adjusting retrieval rank based on POS information, rather than filtering documents out completely. We also propose to attempt a learning-to-rank approach to handle reranking within the POS-constrained query set.
In addition to reranking via other methods, we wish to explore a factor-based MT approach to add POS tags directly to the translation process. This approach may help mitigate some of the noise present in both translation and POS tagging, and help POS-constrained CLIR overall.

Acknowledgement

I would like to thank Professor Dragomir Radev for his assistance and guidance in this project. I would also like to thank Professor Doug Oard and our other collaborators at UMD and Columbia for their helpful discussions.