Data Selection in PSQ System for MATERIAL CLIR

Jonathan Xu and Dragomir Radev

Yale University

LILY Lab

Background

For most languages, there are very few tools available for information retrieval (IR) and machine translation (MT). While some of the more commonly spoken languages have natural language processing (NLP) technologies within static domains, this technology is not easily transferable across languages, domains, and different genres of data. The goal of the MATERIAL project is to enable rapid development of language-independent methods to create systems capable of fulfilling cross-language information retrieval tasks over text and speech data. This entails the investigation on how MT and IR methods can be efficiently developed and applied to domain-specific information needs against multilingual data.

The main development in the MATERIAL system involves the use of Automatic Speech Recognition (ASR), MT, cross-language information retrieval (CLIR) and summarization technologies to create functional end-to-end systems. These systems will take English domain-contextualized queries and produce cross-language English summaries of the query results that describe the relevance of retrieved documents to the original query.

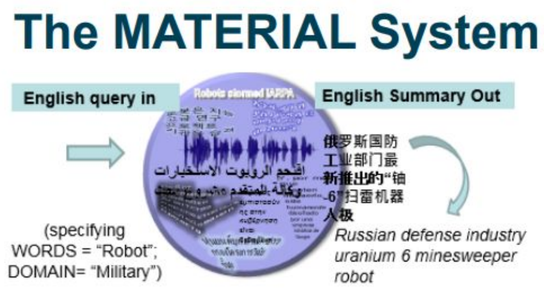
The system that Yale is helping develop for the MATERIAL project is called SCRIPTS and denotes a “fully automatic modular java system that integrates all components into a simple easy-to-use one system.” SCRIPTS combines many underlying components to create an end-to-end system. These components include text processing, summarization, speech processing and cross-language information retrieval.

Data and System

The main goal of this project is to improve CLIR results within SCRIPTS by selectively limiting the training data used for translation. Because of the many different systems used for CLIR, the goal was to improve results in each component individually before implementing it throughout the entire system.

The system we began with was the probabilistic structured query (PSQ) system, which relies on a translation matrix in order to retrieve documents from a query. This sparse matrix provides the probability that words in a foreign language represent terms in english. Once this matrix is provided, the system attempts to translate of foreign documents in order to accomplish information retrieval based off the input query.

There are three main components of data that are used for information retrieval in the PSQ system. The first is a list of queries to retrieve documents for, the second is a list of documents to retrieve from, and the third is a set of translation pairs used to generate the translation matrix. These translation pairs, called the parallel corpus, come in the form of phrases, sentences, or words of the foreign language alongside an official translation in english. There are 20 distinct sources that separate the parallel corpus originating from a variety of online datasets or projects. Through these 20 sources, there are around 10 million different pairs available.



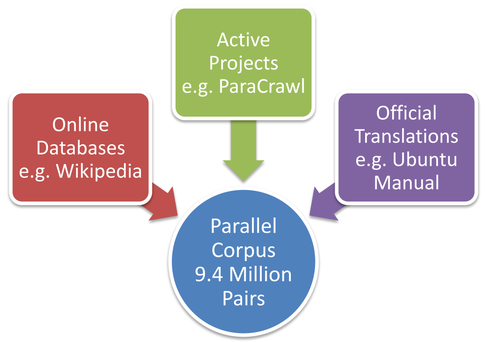


Figure 1. Variety of sources for the parallel corpus

Data Set	Number of Translation Pairs	Map Score
Full Parallel Corpus	9,399,627	0.5649
Best 10,000	176,815	0.5170
Equal Fourths	975,739	0.5353
Random Fourths	2,536,747	0.3877

Figure 2. Map scores from the different selected subsets.

Methods and Results

The main approach to data selection involved taking each translation pair as its own document and using term frequency–inverse document frequency (TFIDF) analysis. TFIDF takes into account both the frequency of the word within a document and its appearance throughout the document set. With this information, it can calculate a weighted value that determines the relevance of any document. Using TFIDF and treating each translation pair as a document allows us to find the most related pairs based purely on the words used in a query. Through this method I was able to derive subsets of the parallel corpus most related to the terms used in the list of queries.

There were two main subsets that I tested, one which took an equal number of translation pairs from each source, and one that took a percentage of pairs from each source. In order to gain a proper comparison, I also tested a subset of the pairs consisting of a randomized selection. The results can be seen in the table below.

Best 10,000 represents the score from choosing the best 10,000 translation pairs from each source. Equal fourths represents taking the best quarter from each source while random fourth represents a data set of approximately one-fourth the size of the full dataset selected randomly. Equal fourths is noticeably smaller than random fourths due to difficulties with calculating tf-idf scores with a particularly large source.

Conclusion

MAP scores represented a weighted comparison between the output of the system using our generated probability matrix and the ideal document set for retrieval. As seen in the graph, both selected subsets performed better than the randomly selected subset. However, they also both performed worse than using the entire parallel corpus as input. While this isn’t the best possible outcome, it's promising that these subsets perform markedly better than random translation pairs of similar size. Perhaps a finer tuning of the number of translation pairs or their distribution across sources will provide an improvement.

While TFIF is great at identifying if related words exist within translation pairs, it does not cover larger phrases or similar, but not identical, words. Because of this, there are many other metrics that can be used to do data selection. The next steps for this project would be to test and/or incorporate other methods to find the optimal recommended subset.

Acknowledgement

I would like to acknowledge Professor Radev and all of the people on the Material Team for their advice and help throughout the project