Extracting Paraphrases from Aligned Corpora

Ali Ibrahim & Boris Katz

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The expressiveness of human language allows people to express the same idea in many different ways; they may use different words to refer to the same entity or employ different phrases to describe the same concept. Thus, an effective information retrieval (IR) and question answering (QA) system must be equipped to handle these variations, both when processing documents and when fielding queries. While there are many resources that help systems deal with single word synonyms, there are few resources for multiple word paraphrases.

Motivation: Synonymy has long been recognized as an important factor in the performance of IR systems. Most research has focused on single-word synonyms; while this approach is useful, it still does not address the issue of paraphrases, or multi-word synonymy. Unlike single word thesauri, multiple-word thesauri are hard to construct by hand. We wish to automatically construct such thesauri from available free text for use in information retrieval.

Previous Work: Barzilay [1] extracted both single and multiple word paraphrases from a sentence-aligned corpus for use in multi-document summarization. Sentence-aligned corpora refer to texts that have been aligned at the sentence level, i.e., a one-to-one correspondence between sentences, where each matching pair expresses the same general ideas or concepts. For example, here is an alignment from two different translations of 20,000 Leagues Under the Sea, Chapter 21:

Ned Land tried the soil with his feet, as if to take possession of it.
Ned Land tested the soil with his foot, as if he were laying claim to it.

Barzilay constructed her aligned corpus from multiple translations of foreign novels. Using this corpus, she co-trained a classifier that could decide whether or not two phrases were paraphrases of each other based on their surrounding context (words occurring to the left and right of the phrases). Barzilay collected 9483 paraphrases with an average precision of 85.5%. However, 70.8% of the paraphrases were single words. In addition, the paraphrases were required to be contiguous.

Lin [3] used a general text corpus to extract paraphrases or inference rules. In his algorithm, paraphrases are represented as matching dependency tree paths between two words. The two words at the ends of a path are considered to be features of that path. For each path, he recorded the list of features that were associated with the path and their respective frequencies. Lin calculated the similarity of two paths by looking at the similarity of their features. This method allowed him to extract paraphrases of moderate length from general corpora. However, the technique is computationally expensive, and furthermore can give misleading results (i.e., paraphrases having the opposite meaning often share the same path).

Approach: We address the problem of extracting multi-word paraphrases by combining the approaches of Barzilay and Lin. By expressing the paraphrases using Lin’s representation and utilizing Barzilay’s corpus alignment techniques, we hope to ameliorate the drawbacks of both approaches. Based on the hypothesis that paths between same words in aligned sentences are semantically equivalent, we can extract paraphrases by scoring the path frequency and context.

Our algorithm consists of four stages (See Figure 1). First, two translation corpora are aligned by a variation we developed of the Gale and Church [2] algorithm. The aligned corpora are then parsed by the Link Parser [5] developed at CMU. The parsed output is then modified to make the links more consistent and semantically meaningful. Finally, paraphrases are generated and scored.

Impact: Using seven translations of three different foreign novels, we obtained 6937 paraphrases. Manually judging a random sample of the results, we found 44% of them to be valid paraphrases. Here are a few interesting
Figure 1: The major stages of our algorithm

Examples:

\[ \text{put} \rightarrow K \rightarrow \text{on} \iff \text{wear} \]

\text{Example: John put on his sweater} \iff \text{John wore his sweater}

\[ \text{rush} \rightarrow K \rightarrow \text{over} \rightarrow K \rightarrow \text{to} \iff \text{run} \rightarrow MV \rightarrow \text{to} \]

\text{Example: And he rushed over to his son, who had just jumped into a heap of lime to whiten his shoes.} 
\iff \text{And he ran to his son, who had just precipitated himself into a heap of lime in order to whiten his boots.}

The paraphrases generated by our algorithm could prove to be useful in several natural language applications; however, we focus on applications in information retrieval. Paraphrases could be used to expand queries or normalize the representation of documents. The Sapere [4] system is a natural candidate for experiments, because it indexes relations, which are essentially the same representation as the paths in our paraphrases.

Future work: There are several avenues that can be further explored to improve the accuracy and coverage of the paraphrases. Foremost, our algorithm’s effectiveness is highly correlated with the amount of aligned data available. In addition to obtaining more translations of foreign novels, we can derive aligned data from newspaper articles and encyclopedia entries describing common events or ideas. While currently inadequate for these types of corpora, our alignment tools can be improved by incorporating discourse considerations and using single-word thesauri. The problem of recognizing the essential elements of a paraphrase is central to increasing the coverage of our algorithm; more work could be done in this respect. Finally, we can also try to extract general morpho-syntactic rules by examining many similar paraphrases.

Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:


