Tools and Techniques for Automatic Bilingual Lexicon Construction

By: Brian Rassier
Faculty Sponsor: Dr. Ted Pedersen

Statement of Problem

The objective of this project is to implement and evaluate techniques that automatically identify word correspondences in parallel texts. This will allow us to automatically construct a bilingual lexicon that is tailored to a particular subject domain and/or pair of languages.

A document and its translation into a second language are said to form a parallel text. Word correspondences are pairs of words (or phrases) that are translations of each other. Thus, the techniques we develop will automatically identify which words and phrases are translations of each other in a pair of texts. This is a challenging problem, since the order in which words and sentences occur in a translation differ considerably from the original text. In addition, it is relatively common for a single word in one language to be translated as a phrase in the other language. Thus, a naive method that assumes that the n^{th} word in the translated text is the translation of the n^{th} word in the original text perform very poorly.

The ability to automatically construct bilingual lexicons is important given the large amount of multilingual text that is now available online. Even casual users of the Internet frequently encounter information in many different languages, and the availability of bilingual lexicons for a wide range of languages and subjects may be useful to users who wish to decide whether or not a document is relevant. Bilingual lexicons can can also be used to achieve simple automatic word by word translation that is sometimes sufficient to allow a human to decide if a text merits higher quality human translation.

Method of Inquiry

We will use techniques from Machine Learning [Mitchell97B] to carry out this investigation. As a part of an Undergraduate Research Assistantship in the Spring of 2002 (funded by Dr. Pedersen's CAREER grant from the National Science Foundation) I have developed a graphical software tool that allows a human expert to manually perform word alignment between text in two languages. The data created with this tool will serve as training data that we will use as input to Machine Learning algorithms to find general rules and patterns for determining how word associations can be found in texts for a given language pair.

Dr. Pedersen has worked extensively with decision tree learning [Pedersen01b] so this will be our main focus. Among the advantages of decision trees is that there are a number of efficient and well understood algorithms available that allow them to be learned automatically from training data (e.g., [Quinlan86], [Weka]). Also, decision trees are easily interpreted by humans and can be converted into sets of rules that provide insight into the problem.

Our new approach will be evaluated relative to a well known method for finding word correspondences in text, the k-vec algorithm [FungC94].

Expected Results

This work will assess whether or not decision tree learners are an appropriate tool for automatically finding word correspondences in parallel text. Most previous work in finding word correspondences in parallel text has relied on classical pattern recognition techniques (e.g., [FungC94), [Melamed01]) or dynamic programming (e.g., [GaleC93]). Decision trees have not been previously applied to this problem due to the difficulty of obtaining sufficient amounts of training data. However, the tool I developed during Research Assistantship has made it relatively easy to create such data, thereby making it possible to investigate the use of decision trees.

Time Table

This work will be carried out in the summer of 2002. I plan to work 13 hours a week for 10 weeks in June, July, and August.

- Weeks 1 \& 2 (26 hours)

  Develop software tools to convert text into form suitable for decision tree learning.

- Weeks 3 \& 4 (26 hours)

  Carry out initial experiments with decision tree learning, using standard parameter settings.

- Weeks 5 \& 6 (26 hours)

  Evaluation of first set of results. In particular, compare with results from the k-vec algorithm.

- Weeks 7 \& 8 (26 hours)

  Second set of experiments, with refined approach to learning procedure based on first round of work.

- Weeks 9 \& 10 (26 hours)

  Prepare final report and/or poster discussing results.

Budget

We are requesting a budget of $1700. This includes a $1400 stipend to be paid in the summer of 2002, as described above. It also includes $300 for General Operating Supplies, which will be used to purchase the following books (prices are from amazon.com as of February 26):

1. Flexible Pattern Matching in Strings, by Gonzalo Navarro, \$50
2. Data Mining, by Ian Witten, \$50
3. Managing Gigabytes, Ian Witten, \$55
4. The Theory and Practice of Sequence Comparison, by David Sankoff, $50
5. Synonymy and Semantic Classification, Karen Sparck Jones, \$60
6. Programming Perl, Larry Wall, \$35

Relationship to Faculty Sponsor's Research

Dr. Pedersen does research in natural language processing, in particular developing methods that automatically resolve the meanings of ambiguous words in text. Work in automatic methods of translation and bilingual lexicon construction are also of interest to him, since the different meanings of a word in one language are often translated as completely different words in a second language.

Dr. Pedersen has had a Master's student working with the previously mentioned k-vec algorithm, and has considerable software infrastructure already available for performing evaluations of bilingual lexicons.

Educational Objectives

I have studied Spanish an have an interest in language. I am particularly Iinterested in how the World Wide Web has evolved as a multilingual resource. In the future I would like to work as a programmer, and I believe that experience with challenging language processing applications such as this will be very useful.