Tools and Techniques for Automatic Bilingual Lexicon Construction
UROP (Summer 2002) Final Report

By: Brian Rassier

Faculty Sponsor: Dr. Ted Pedersen

September 6, 2002

My UROP for the summer of 2002 dealt with automatically constructing bilingual dictionaries from large amounts of parallel text. A document and its translation into a second language are said to form parallel text. Word correspondences are pairs of words that are translations of each other, so developing methods that automatically discover word correspondences in parallel text are the foundation of automatic methods to construct bilingual dictionaries.

My initial objective in this project was to implement and evaluate techniques that automatically identify such correspondences in parallel text. I had planned to explore supervised learning techniques such as decision tree learning to carry out this objective. Supervised learning algorithms are very powerful, but they require the availability of examples from which to learn. This is where I encountered a serious problem. The types of examples these methods require are parallel texts that have been manually aligned. This requires that a human expert go through the text manually and decide which words are translations of each other, and then indicate this somehow in a form that a computer program can utilize. As of this date (September 2002) there are very small amounts of such manually created data available. The Blinker data from New York University consists of 500 translated verses from French and English versions of the Bible that have been manually aligned. There is also 250 sentences from the French and English proceedings of the Canadian government that have been manually aligned by a research group in Germany. I quickly realized that this was not a significant enough quantity of data to carry out my initial objectives.

I investigated why there was so little data available to the research community, and concluded that one factor was the lack of good tools for human beings to use in identifying which words are translations of each other in a text. The process of manually aligning parallel texts is time consuming under the best of circumstances, and without a good computer based tool it is nearly impossible. At present there are no publicly available tools to aid in the creation of this type of data, so Dr. Pedersen and I decided that we could make a large contribution to the research community by creating a tool that provides a graphical user interface to human experts engaged in the creation of word aligned text.

The tool is named Alpaco (Aligner for parallel corpora), and it enables users to create a set of word correspondences in parallel texts. The aligned data is stored in such a way that it is easy for a computer program to use it as input. Thus, Alpaco fills an essential

role in the automatic creation of bilingual lexicons. It will make it possible to quickly increase the amount of training data available for supervised learning algorithms, and this will encourage both the development of increased quantities of data and will also encourage researchers to consider such approaches. In the past these methods have not been widely pursued, presumably due to the difficulties in creating sufficient quantities of example data.

Bilingual lexicons are important because they play a large role in computational approaches to translation of human languages. This area is sometimes called Machine Translation. The lexicons can be used to automatically translate on a word-by-word basis, and in some cases phrase-by-phrase. There are instances in all languages where a phrase may not make sense when broken down word-by-word into another language. Alpaco allows for this and lets users create correspondences on a phrasal level. This will help to create more accurate bilingual lexicons, and more accurate translation will follow.

One area where considerable research is possible is in how to represent the word-by-word correspondences that a user creates. We settled on breaking up the two parallel texts into sentence-aligned text. Alpaco then gives each word a numerical identification. Pairs of these numerical identifications are kept, which indicates a connection between the two words. The format that word-by-word correspondences are recorded may be an area for future work. There have recently been breakthroughs in creating a standard for representing these correspondences. Alpaco could be changed to represent its alignments in this standard.

Alpaco is still under construction and has not yet been made publicly available to the research community. However, a preliminary version is available on my home page: http://www.d.umn.edu/~rass0028. Additional technical documentation and user instruction is available at: http://www.d.umn.edu/~rass0028/README.txt. There are some additional features that need to be incorporated and further stress testing must be carried out. Dr. Pedersen has decided that this work is of sufficient potential impact that he will continue to support me as a 10-hour per week Undergraduate Research Assistant in the Fall of 2002 as a part of his CAREER grant from the National Science Foundation. We will release Alpaco under the Gnu Copy Left, which means that the source code is freely available and can be incorporated into other projects as long as we are given appropriate credit for our work. Alpaco will be made available in the Fall of 2002, and we have already had expressions of interest from research groups at New York University, the University of Maryland, the University of Toronto, and the University of Southern California. We look forward to getting their feedback, as well as that of other users. Dr. Pedersen has released other software in this manner, and it has always been well received so I am hopeful the same will be true for Alpaco.

During my UROP I also did research on web-based programming. Another area for future work with Alpaco could be to transfer its functionality to a web-based program. This could be very beneficial because there is a large amount of multi-lingual parallel text on the Internet. Users around the world could then voluntarily align small amounts of

text, and in so doing create a huge repository of valuable data for researchers in Machine Translation.

Of course now that Alpaco is nearly finished, we will be able to easily create data suitable for the original goals of this project. That work will proceed, and Dr. Pedersen is optimistic that the availability of additional data will result in significant advances to automatically constructing bilingual lexicons via supervised learning.

While Dr. Pedersen will support the completion of Alpaco with his CAREER grant, their remain additional issues that can be pursued via other programs, possibly even future UROPs. For example, developing a web based tool that allows users from around the world to align text in various languages could be a significant source of data for relatively understudied languages. Also, carrying out the work in supervised learning with the data created by Alpaco remains an interesting and important area of work.

Despite the change of direction, we accomplished many of our goals, and I think the UROP was very successful. I learned many things that I would otherwise not have had the chance to learn. I learned computer languages that I wouldn't have learned in classes, gained experience in developing software, and learned more about the field of natural language processing. All of these experiences could easily transfer to my future career. The work I did also showed me how much I enjoy developing software. This is helpful because it gives me an idea of some careers that I would enjoy when I graduate this spring. I also learned that research must be flexible, and that sometimes when looking at an interesting problem you realize that you simply can’t proceed any further until certain infrastructure is developed.

My faculty sponsor, Dr. Pedersen, was also a huge help with the project. He had many great ideas, and was the inspiration for the bulk of this project. He contributed a wealth of knowledge, insight, and suggestions. The project could not have been done without him.