Building Resources for Languages with Scarce Resources

By: Brian Rassier

Faculty Sponsor: Dr. Ted Pedersen

October 1, 2002

Statement of Problem

The objective of this UROP is to develop a Web interface that will allow for the construction of a large collection of multilingual parallel text with the help of Web users from around the world. This is an important task because parallel text (i.e., a text in one language and it's translation in another) is a useful source of information for automated approaches to translation of human languages. We are particularly interested in collecting text from languages that are not yet widely studied in natural language processing or computational linguistics. While there is quite a lot of data available for languages such as English, French, and Japanese, there are very limited amounts of parallel text available for most of the other languages of the world.

Recent events have demonstrated the need for being able to quickly develop systems that can process and translate languages where there is limited parallel text available such as Arabic, Farsi, and Pashto. Moreover, as noted in a recent article in the Scientific American [1], a significant percentage of today's 7,200 different spoken languages are close to extinction, and therefore there a great need for sustained conservation efforts in relation to these languages. We believe the encouraging native speakers of these languages to become involved in the creation of parallel text will make it possible to develop translation systems in those languages more quickly, and also promote the continued use of these languages.

The data created will have two uses. First, it will make it possible for speakers of languages to make news and information from their country and culture available to those who do not know the language. In addition, we will store the parallel text in a form that is suitable for use by researchers in natural language processing and computational linguistics who are interested in developing tools and techniques suitable for parallel texts.

Method of Inquiry

Parallel Text Collection

We plan to study the issues related to the collection of parallel texts over the Web. Specifically, we will be dealing with two types of problems. First, we need to motivate the users to participate in this process of creating parallel data. The data collection site will be set as a news site where users can contribute translations of the daily news in their own language. This subject is likely to attract the interest of a high number of users. We also plan to identify sources of bilingual users whose native languages are not widely

used. Second, since this is a non-conventional approach that allows *every* Web user to contribute his/her knowledge to our data repository, we must develop the means for verifying the correctness of the translations. After we have developed the interface for users to contribute translation, the following topics will be investigated (as time permits).

- Language identification: create tools that are able to identify the language used in a given text, for a large range of languages.

- Translation identification: create tools that are able to identify the equivalence between two translated texts. Initially, this will involve language identification, length and cognates validation. As the collection of parallel data will grow, we will investigate means of using previously translated texts to identify new translations equivalence.

- Sentence and word boundary detection: in languages like English words are usually delimited by spaces and sentences by a small set of punctuation. This is not true of all languages, and the issues of identifying where words and sentences begin and end is considerably more complex.

- Comparative evaluation: the contributed translations will be compared with the output of publicly available machine translation tools and the judgment of other translators.

Tools and Resources for Languages with Scarce Resources

As a part of our Web based collection of translated text, we plan to investigate the derivation of resources from *multilingual* texts. Multilingual texts are those that have been translated into more than one other language. These are even less common than bilingual texts, and previous research in machine translation and related areas was generally limited to bilingual data. The only notable exception is a study on sentence alignment performed in 1999 at the University of Montreal [2], which demonstrated that even for a very small text, the use of three languages brings significant improvement in the alignment quality as compared to the use of only two languages. No other well-known multilingual studies were performed.

While the use of multilingual parallel data is highly likely to bring significant improvements in the performance of machine translation, the lack of prior research in this area is due to the lack of multilingual data. We plan to overcome this drawback by collecting large amounts of multilingual parallel data over the Web. Specifically, we will conduct research to investigate the impact of multilingual parallel corpora (as compared to bilingual parallel corpora) on:

- Word alignment: the basic problem after parallel text has been collected is to determine which words are translations of each other. This information can be used to develop translation systems and also to create bilingual lexicons and dictionaries. This is recognized as a rather difficult problem for parallel text in two languages, but

we believe that solutions based on multilingual text (translations in two or more languages of a given text) will be more successful.

- Multilingual dictionary derivation: lists of translated terms are a natural byproduct of word alignment. In the case of multilingual text, we will develop methods to create trilingual dictionaries.

Dr. Pedersen has at least one graduate student working on these two problems, and that student will benefit from the data collected by this UROP project.

Expected Results

This work will create a Web site where users from around the world can volunteer to translate articles in their native language into another language. This will help to spread information, especially we hope from languages and cultures that are not currently well represented on the Web. The data collected by this effort will be used to determine how multilingual parallel texts can be used to improve the precision of word alignments and language translation. All of the data we collect will be made freely available to researchers working on these problems.

References

[1] W.W. Gibbs, Saving Dying Languages, Scientific American, Aug.2002
[2] Simard, M., EMNLP 1999, Text-Translation Alignment: Three Languages Are Better Than Two.

Time Table

The project will be carried out in the spring semester of 2003. I plan to work 10 hours per week for 13 weeks in January, February, March and April.

- Weeks 1-5 (50 hours). Develop a Web-based translation program to collect multilingual parallel texts.
- Weeks 6 & 7 (20 hours). Carry out initial experiments with language identification and sentence boundary detection in multiple languages.
- Weeks 8-10 (30 hours). Carry out experiments with word alignments and parallel lexicons derived through multilingual parallel texts.
- Weeks 11-13 (30 hours). Second set of experiments based on the multilingual data collected from the Web-based translation program. Prepare final report discussing the results of the project.

Budget

We are requesting a budget of $1676. This includes a $1400 stipend to be paid over the spring semester of 2003 as described above. It also includes $276 for General Operating Supplies, which will be used to purchase the following books (prices are from amazon.com as of October 1, 2002):

1. Weaving a Website: Programming in HTML, Java Script, Perl and Java, by Susan Anderson-Freed, $60

2. Programming Language Pragmatics, by Michael L. Scott, $74

3. Lexical Semantics and Knowledge Representation in Multilingual Text Generation: The Kluwer International Series in Engineering and Computer Science, by Manfred Stede, $142

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. I worked with Dr. Pedersen in the spring of 2002 as an Undergraduate Research Assistant (funded by Dr. Pedersen's NSF CAREER grant) and during the summer of 2002 via a UROP. The work carried out there has resulted in the creation of a freely available program that allows a human expert to manually word align parallel text. This software is currently being used by one of Dr. Pedersen's graduate students to create word-aligned data suitable for automatically creating a dictionary from parallel text. It will also be made freely available for use by researchers around the world.

Educational Objectives

I took a class in natural language processing in the spring of 2002 which sparked my interest in the field. This project will extend my understanding of parallel texts, word correspondences, and language translation. I am interested to see how the multilingual parallel texts will affect the results of these areas. I also think the collection of these multilingual parallel texts will benefit the field of natural language processing in a number of areas (e.g. language translation, lexicon construction). I will be graduating this spring, and am planning on starting a career in computer programming. I think that this experience in the challenging field of language processing will be very beneficial to my career.