Suggestions for Collaboration between Avenue and Catanal

Lori Levin, Alon Lavie, Jaime Carbonell, Alex Waibel, Ralf Brown, John Lafferty
Language Technologies Institute
Carnegie Mellon University
lsl@cs.cmu.edu

1 Introduction

Machine Translation (MT) systems have been built for a variety of language pairs. A quick glance at the Compendium of MT Software suggests that English and Japanese are the most popular languages for commercial MT systems. Major European languages come next, followed by a few minor European languages (e.g., Basque and Slovenian), and some Middle Eastern languages. Finally the compendium shows less than a handful of MT systems for minor indigenous languages in Africa, North America, and the former Soviet Union. The reason for the bias toward the languages of major trade partners is largely economic: MT systems are expensive to develop. Some commercial systems represent a person-century of work.

The absence of MT research on indigenous languages is a missed opportunity socially and a tragedy scientifically. From the scientific point of view MT systems are being developed without taking into account all types of languages. For example, we do not know of an MT system for a polysynthetic or incorporating language, and we do not know whether all current MT techniques are suitable for those types of languages. This is a flaw in research that is supposed to be applicable to all languages. The missed social opportunity is for speakers of indigenous languages to participate in government, education, health care, and the internet without having to give up their languages.

This white paper sketches some plans for collaboration of the Catanal project with Carnegie Mellon's Avenue project. The suggested role of Avenue is to provide a machine translation program, whereas the suggested role of Catanal is to assist the native community in finding uses for MT in education and language preservation, to design front end applications using MT as a back end, and to install MT-based technology in the communities.

First we will give a brief history of Avenue (which was originally called Nice) then describe some basic issues in MT design, and finally describe potential research directions.

2 History of the NICE/AVENUE Project

The Nice (Native language interpretation and communication environment) project was originally created Gary Strong (NSF) in response to a series of NSF/OAS workshops on Western Hemisphere collaboration. Machine Translation was proposed as a step toward digital government, closing the digital divide, and better communication between governments and citizens. However, in order to close the digital divide, MT needed to be made available to people who speak neither Spanish nor English. Jaime Carbonell, Director of the Language Technologies Institute, attended some of the workshops. Since LTI had already been doing research on MT for minor languages, the Nice project was created and funded at LTI.

The original funding came from DARPA during the time that Gary Strong was a program manager there.

Prior to the Nice project, Peter Wilkniss had organized a workshop for Catanal, Computer Aided Translation of Alaska Native Languages. Lori Levin attended that workshop, and used it as a model for Nice, the most important points being that the project is a partnership of scientists and native people, and that the native people design the project for the needs of their communities.

Nice then formed a partnership with the Institute for Indigenous Studies at the Universidad de la Frontera, in Temuco, Chile. This partnership has been in place for about 1.5 years. The team in Chile is recording, transcribing, and translating data in Mapudungun (the language of the Mapuche people). The data will be used as training data for MT research. So far 80 hours of speech have been recorded and transcribed, and 200K words of paper text have been typed into electronic form. The choice of data and issues of orthography in the control of the team in Chile. The data is recorded and transcribed in Chile under the supervision of Mapuche linguists and is freely available to the Mapuche community for whatever uses they choose.

LTI researchers would also like to pursue a partnership with Catanal and the Iñupiat community in Barrow, Alaska. In addition the original Catanal workshop, two meetings have taken place. A brief meeting at the Alaska Native Educators Conference in Anchorage and a two-day meeting in Barrow.

Recently, because of changes in program managers, the work of the Nice project has been taken over by two new projects at the LTI: A DARPA project called MilliRADD (Rapidly Adaptable Data Driven MT for small amounts of training data), and an NSF ITR Project called Avenue. Research on indigenous languages is continuing under Avenue. MilliRADD is focusing on Chinese and Arabic so as to have a common ground for evaluation in comparison to other DARPA funded MT projects.

3 Some Basic Issues in Machine Translation

An ideal MT system would produce high quality translations of material in every style of writing or speaking on every topic, and would require only a few weeks to adapt to a new language. In actual practice, each MT system achieves some, but not all of these goals. For example, one MT system produces high quality translations of technical manuals, but requires the input language to be closely constrained, and takes years to develop. This is useful for an industry that needs to disseminate clear technical material to its customers. Another MT system can be adapted cheaply to new languages in a matter of weeks, but provides a rough translation that is just good enough to give a gist of the meaning to an intelligence analyst. Spoken language translation systems for face-to-face conversation can handle disfluent spontaneously produced speech, but are limited to specific topics (like travel arrangements), and produce medium quality translations that can be negotiated and repaired by the participants in the conversation.

In short, there are many types of MT, and each MT project has to be designed to match what it will be used for. For a collaboration between Catanal and Avenue, the primary uses might be translation of educational material (e.g., science, math, and history) into

Iñupiaq for use in immersion classes, or development of lessons for teaching the Iñupiaq language itself. In the following sections, we review the pros and cons of different types of MT for the Northern Alaska setting.

4 Human Engineered vs Automatically Learned MT

4.1 Human Engineered MT

Human Engineered MT is a tried and true method. It produces the highest quality output. To produce a human-engineered system, linguists and computational linguists write rules for the source and target languages. Following are the pros and cons:

- **Advantages of Human Engineered MT**
  - High quality output: This is important for a community that wants to really use MT, and not just participate in risky research.
  - Can start right away: Human linguists can start writing rules without waiting for lengthy data collection procedures.
  - Human-readable rules: The system consists of coded grammar rules that can be read and modified by humans.
- **Disadvantages of Human Engineered MT**
  - Long development time: In the end, the MT system produces high quality output, but it may take up to two years for there to be broad coverage.
  - Known method/Not new science: Work on human-engineered MT would probably not be funded by NSF's CISE or linguistics programs.
  - Extensive human expertise required: Building a human engineered MT system requires humans that are fluent in both languages and who are trained in computational linguistics.

4.2 Automatically Learned MT

There are many methods for developing MT systems without human linguists. These methods consist of computer programs that learn translation correspondences from a corpus of previously translated texts.

- **Advantages of Automatically Learned MT:**
  - Participate in cutting edge research: Automatic methods have been designed by computer scientists who are not familiar with language typology and the range of differences between languages. Speakers of polysynthetic languages can participate in this research and ensure that the methods that are developed are applicable to all language types.

  - Can be built quickly (after data collection): After a large enough database has been collected (see below), the first version of the MT system can be running in a matter of weeks.
  - Division of Labor: A computer scientist can build the MT system by feeding it previously translated texts without knowing anything about the language. Native speakers provide the texts and do not have to know computer science or computational linguistics.
  - Coverage: An automatically learned system can produce some kind of output for any input. A human-engineered system only covers material that the humans have had time to cover.

- **Disadvantages of Automatically Learned MT:**
  - Data requirements: Algorithms for automatically learning translation correspondences must be trained on corpora that are over one million words long. Some methods require ten million words. The reason is that the program needs to see many different ways of saying the same thing in many different contexts and also needs a statistically significant number of examples to work with. If such a corpus doesn't already exist, it can take more than six months to create it by gathering and typing material. Current research has made some progress on reducing the data requirements.
  - Quality: The state of the art in automatically learned MT is that the quality is not very high. DARPA finds it useful because it may be just readable enough to give a gist of the meaning, and then the document can be handed off to a language expert for more careful translation. Maybe it would be useful for web browsing.
  - Not human-readable: The translation "rules" of an automatically learned MT program are actually a collection of statistics that cannot be understood by humans. Of course, a human can change the "rules" system by changing the learning algorithm and the learning data, and then letting the system learn again.

4.3 Hybrid Systems

A variety of methods are available for mixing human-engineered and rule-based systems. One method is multi-engine MT. Under this method, a human-engineered and an automatically-learned system both translate the same material, and their outputs are combined statistically in a way that uses the best from both methods.

5 Recommendations and Estimates of Effort

- Carnegie Mellon is already investigating several methods for automatically learned MT. We would be very happy to have the Iñupiat community participate with us on this research, especially since we would like to test our methods on polysynthetic languages. Carnegie Mellon can do this work under its current AVENUE funding.

- In order for the Iñupiat community to participate in the research on automatically learned methods, the community would have to gather up its lexicons/dictionaries and texts. This might be up to six months work. AVENUE could supply some seed funding to get it started, but this data collection effort would then require more funding – maybe two people for six months with some supervision by knowledgeable community leaders. Collection of spoken data for speech recognition would take an additional few months. The data collection effort would be of general use to the Iñupiat community (e.g., reviewing elders' conferences) beyond its use for machine translation.

- There should be a human engineered component to the project. If the quality of automatically learned methods is not adequate, Iñupiat people may feel that their time and money has been wasted. They need an MT system, not just risky research. However, the human-engineered component cannot be funded by CISE because it is a known technique and not new research. This work would need a significant amount of funding: at least one full time Iñupiat speaker for two years, part time supervision by community leaders with knowledge of linguistics, one graduate student at Carnegie Mellon for two years, and faculty supervision at Carnegie Mellon.

- Multi-engine integration of human-engineered and automatically learned systems. Carnegie Mellon can do this under AVENUE funding.

- Front-end application: After a machine translation problem is like a bare engine for a vehicle. The rest of the vehicle needs to be designed and deployed: for example, who will use it and for what purpose, what kind of user interface will it have, if it will be used in education, which part of the curriculum will it fit into and what kind of educational lesson plans go with it. The CATANAL project and the Iñupiat community will carry out this part of the work, which will also require funding.