R&D for Commercial MT

Laurie Gerber

SYSTRAN Software, Inc.
7855 Fay Avenue, Suite 300
La Jolla, CA 92037
(619) 459-6700
lgerber@systransoft.com

Abstract

MT research in the commercial environment tends to be conservative, and to introduce change gradually, both because of limited funds, and the need to quickly turn innovations into product features. However, there are a number of challenges and opportunities that could make commercial research a much more dynamic environment for advancement of the field as a whole.

1. Commercial Enterprise

There is an assumption in for-profit enterprises, especially those that are not yet hugely fat with surplus revenue, that all activities undertaken as part of the business must serve the bottom line. Unlike academic researchers whose minds are filled with methodological, theoretical (and funding) issues, the commercial researcher often must focus on ways to increase efficiency. The category of things-to-be-researched may be defined ad hoc, as information is needed. This means that while areas of research include experimentation with, or introduction of, new methods or tools, the research arena may stretch from requirements gathering, to improved tools for linguistic programmers, to patent and copyright issues, to keeping abreast of the competition. For the present discussion, I will set aside the miscellany of the research department and focus on research that targets the methodological and theoretical evolution of commercial MT systems. In the final section, I will address the balance of research efforts.

2. The Nature of Commercial R&D

The relative leanness of the MT industry means that, in general, research goals are subordinate to production goals when priorities are established, and that research tends to focus on implementation, rather than experimentation. Ideas and methods developed elsewhere must be carefully considered and reviewed to see if they are suitable to incorporate into a production system. It is rare for a commercial MT developer to have staff members who can focus exclusively on research concerns, and who have no production-related responsibilities.

By way of contrast with academic research, there are advantages and disadvantages. Commercial researchers have the satisfaction of knowing that their work will be incorporated in a working product that may soon be enjoyed by many users. However, we face the limitation that our research projects must fit the product profile, and must have a strong potential to be

efficiently scaled up for production use. We generally do not have the freedom to explore areas that do not address user concerns. In contrast to academic researchers who can approach research with something like "scientific method" just to test the validity of assumptions or ideas, the topics of commercial research tend to be items that already have a high probability of success. However, in exchange for some of the freedom that academic researchers appear to have, we get the security of a defined revenue stream, which gives us some greater notion of our future. Finally, commercial research doesn't face the same pressure from "fashion". Since publication is not a primary objective, we are not ashamed to use methods that are considered old or unfashionable if they meet our needs.

3. Leverage and Baggage - the Dynamics of Progress

A commercial system with several languages in production has significant advantages when undertaking research. Large lexicons are already available, some kind of text corpora have probably been accumulated for testing, and all the tools used the course of daily development activities are already on hand. All of these mean that when the commercial researcher does have a chance to research linguistic or methodological issues, time does not need to be spent in setting up infrastructure. The other side of this, however, is that once a new technique or system modification is to be introduced, it generally has to work with all languages. If an innovation doesn't include backward compatibility to deal with the idiosyncrasies of older language pairs, a developer faces the headache of maintaining several generations of products based on different methodologies. In the case of SYSTRAN, we place a priority on bringing all language pair systems forward together - all 30 of them!

4. Trickle-down Research

The state of the art doesn't advance very far or fast without help. The (D)ARPA-funded Human Language Technology project of the late 80's and early 90's introduced a number of innovations to the field. The HLT project was a seminal event for both commercial and academic MT research. In their original form, example based, and statistically based methods seemed inaccessible and even incompatible with existing systems. But the innovations that emerged have gradually evolved from their "theoretically pure" initial incarnations, to a wide variety of methods and tools that are now generally available to apply to MT and NLP tasks. As academic researchers publish the results of their work and generously make some of their tools freely available, the whole state of the art can advance in the wake of those innovations, benefiting academic and commercial researchers alike.

Although the early 90's when many of these new findings were being published were a very exciting time for the state of the art, now is the fertile time for advancement of the state of the practice.

5. Challenges

I would like to prognosticate about the brilliant future of research for commercial MT. The challenges described above, such as the time and cost of modernizing existing systems, and the limitation of commercial research to conservative and production oriented projects, can both be

addressed with money - whether it comes in the form of robust profitability, or investment from the outside. The availability of such money will depend on how effective MT developers are in finding success stories for their products, and how promising MT looks to investors and policy-makers.

But there are other challenges, that I will categorize as "cultural", which hinder advancement of the state of the practice in commercial MT. I am again drawing a distinction between the "state of the art" found in (typically small-scale) experimental systems, and the "state of the practice" as seen in existing commercial MT systems, which are largely based on 10-20 year-old technology.

- Limited access to "public" resources: Access to public resources is unfairly restricted for commercial groups. Corpora or lexical resources for sale often have a price to commercial groups that is 10 times the price at which the same resources are made available to academic research groups, effectively putting them out of the reach of commercial MT developers. Many of the resources also come with the restriction that they may only be used for research purposes, and must not be used in the development of any commercial products. This seems particularly unfair when academic groups begin to commercialize their systems and compete for development contracts.
- Failure to integrate trained computational linguists into the commercial MT development environment: One of the mysteries of commercial MT is why it employs so few academically trained computational linguists. From my own experience and anecdotes from other developers, I think there are a couple of reasons. One is the concern of both parties over ideological incompatibility - will the newly trained computational linguist be willing to work with the existing system? The second is the difficulty of making the transition from an environment where the priority is on discovery and publication, to an environment where delivery of reliable products is the priority, with all the accompanying drudgery!
- Few partnerships between academic groups and commercial developers: Recently SYSTRAN Software, Inc. has embarked on a very successful, small-scale partnership with the AI group at the Information Sciences Institute at USC. It has been an ideal project because we were confident of success, and of the ability to integrate the results into the production system. However, it is a challenge to find such projects. It is not easy to find a good fit between researcher expertise and commercial developer need. It generally requires that researchers set aside personal or institutional research goals and pure theory to allow integration into the messy world of production MT. On the commercial developer's side, it is necessary to identify places in the system's architecture where hybridization will be feasible.
- Little usability research: An area of research that is surprisingly neglected is that of user behavior. Maybe this is a holdover from the assumption that linguistic and technical issues are the "big hard problems" that still merit the bulk of research dollars. Knowledge about how to fit into the translation environment is relatively scarce. Developers of translator workstations address these issues for professional, full-time translators, but this is only one subset of MT users. A one-interface-fits-all approach is unlikely to optimally serve the full range of MT users: the business person who uses MT for correspondence or email; the publication department translator (who fits the professional translator model); the scientist or analyst doing research on foreign publications; and the international web browser whose interest may be casual or serious.

6. Opportunities

Commercial MT research holds tremendous potential for improving the quality and usefulness of translation products. The challenges mentioned above are also the areas of greatest opportunity.

- What if the research community relaxed the "research only" requirement, and priced resources based on the anticipated number of users? Or the size of the company buying them instead of the black and white commercial vs. academic pricing?
- Well-developed dictionaries are essential to the construction of high-quality rule-based and knowledge-based MT systems. Commercial MT developers have them. The need to develop dictionaries, sometimes from scratch, is a huge consumer of research money, often delaying or diminishing the work that can be done on the real topic of research. Well designed partnerships or even barter relationships could make those lexical resources available for research, give researchers exposure to the production environment, and facilitate integration of new technology into existing commercial systems.
- Well-designed usability and implementation studies are sorely needed. MT has got to prove its worth in practice. Progress toward FAHQMT has proven painfully slow, yet we can't really abandon it as an ideal and a goal. In the meantime, MT needs to earn its keep, and the best way it through good implementations.

References

Farwell, D. 1996. Observations Concerning Next Steps in MT Research,

Murgida, J. 1996. Ten Years' Experience Talks about the Future! In *Proceedings of the Second Conference of the Association for MT in the Americas.* October 2-5, 1996. Montreal, Quebec, Canada.

O'Hara, J. 1996. Unpublished comments as panelist, "Next Steps in R&D". The Second Conference of the Association for MT in the Americas. October 2-5, 1996. Montreal, Quebec, Canada.