[*Language Industry Monitor* 12, November-December 1992]

TMI ’92: A Second Opinion

Having attended the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI ’92) that is the subject of Tony Whitecomb’s front-page article in the September-October issue of Language Industry Monitor, I am concerned that, as a result of this article, readers will be left with an inaccurate perception of the substantive issues discussed at the conference.

In the opening paragraph, Whitecomb equates “rationalist” with “rule-based” and “empiricist” with “example-based, statistical, or connectionist.” In open discussion at the conference, these absolute associations were acknowledged as misleading, although the organizers did stage a lighthearted “medieval” debate where rationalists, who operate solely by reason and ignore the data, were pitted against empiricists, who scorn anything but observable facts. In real life, there are many who have reasoned at length about the effective application of statistics to the MT problem, and there are also many who have built functional rule-based MT systems by examining a good deal of data.

The value of inspecting large amounts of real text in the development of machine translation systems has been clear for many years to those who have sought to develop real, commercially viable systems. Statistical and example-based methods are an addition welcomed by most researchers in the analysis of and extraction of information from huge text corpora, heretofore considered intractable. But the usefulness of certain linguistic generalizations, often stated as rules, is equally clear, even now to those who have been characterized as “empirical extremists.” Nowhere was this more evident than in Peter Brown’s conference presentation, in which he related how the IBM statistics-based group had begun to incorporate traditional linguistic information at certain levels (e.g., morphological) in their system in order to improve its accuracy.

A hybrid approach

The major point of this article, indeed, its headline, should have been that a combination of the various methods discussed at the conference holds the greatest promise for improved MT systems. Instead, empiricist adulation dwarfed the one sentence toward the end of the article where Whitecomb admits that “Practically all active conference participants agreed that the most likely and promising approach to pursue in the future is a hybrid approach based on example-based, statistics- oriented, and corpora-supported work which is backed by generalized syntactic, lexical, or semantic knowledge.” It is clear that this “generalized ... knowledge” is exactly the knowledge embodied in many of the rules of current, real-world, rule-based MT systems.

I, for one, do not believe that this conference will mark, as stated in the article’s conclusion, “a transition from one fundamental paradigm to another,” but rather that it will be viewed as the point at which the value of statistical, example-based, and related methods for handling certain MT tasks became widely accepted. Furthermore, the value of the tremendous amount of work invested in discovering and implementing many linguistic generalizations found in current rule-based systems will continue to be validated, not relegated to the “scrapheap” as Mercer is quoted as suggesting. The ultimate challenge of building truly viable MT systems will, in the end, force us to use the best that all of these methods have to offer, and it was clear that a majority of the researchers at TMI’92 shared this viewpoint.

- Steve Richardson
Senior Researcher, Natural Language Processing
Microsoft Corporation

Steve Richardson recently co-edited, together with Karen Jensen and George Heidorn, Natural Language Processing: The PLNLP Approach, (*Kluwer* 1992), *a collection of papers detailing their decade of research at IBM*.