Why MT Systems Are Still Not Widely Used?

Keh-Yih SU and Jing-Shin CHANG

Department of Electrical Engineering
National Tsing-Hua University
Hsinchu, Taiwan 30043, ROC
email: kysu@ee.nthu.edu.tw, shin@ee.nthu.edu.tw

1 Problems with MT research and development

Prof. Somers started his topics by raising the question: "Where will MT be in the next 20 years?" As his observations indicate: after years of development of various MT technologies, the second generation MT systems are not much better than the first generation systems, and many commercialized systems are still of the first generation architecture in essence. It is thus interesting to ask "WHY."

Prof. Somers attributes the major problems to the architecture of the second generation MT, where the stratificational paradigm was adopted. The stratificational paradigm results in inappropriate intermediate representations and inappropriate transfer and generation processes that are tightly bound to the source text. Structure preserving translation is common for essentially all such systems. The designers also tend to assume that source text contains enough information to permit translation under such an architecture. As such, the system designers tend to enhance the performance at the source side, including incorporating better semantic theories, dealing with several sentences, accounting the contextual knowledge, and so on.

We generally agree with the observation that stratificational analysis tends to introduce some undesirable effects as mentioned by Prof. Somers. However, we believe that these effects are not the necessary consequences of a stratificational MT architecture. Technical and non-technical problems both contribute to the problems in MT development. So we would like to review the problems from a more general point of view by raising the following questions:

- Why MT systems are still not widely used?
- Why MT methodologies in the current generation are not as good as expected?
  (What's wrong with the second generation MTs?)
- Why MT research advances so slowly?

The first question is related to the different expectations of the customers and the designers. The second question is more interesting to MT researchers

because more technical designing issues are involved. The final point is closely related to the testing environment of the MT community. We raise these problems together, because it is not the technical problem alone that affects the MT advance in the history of MT development.

2 Why MT systems are still not widely use?

Since 1989, the **BehaviorTran** (formerly the **ArchTran**) English-Chinese MT system (Chen, 1991) was put into operation in a translation service center. We found that a customer's view is quite different from a designer's view. The differences might partially explain why MT systems are not widely used today and why a customer is willing to accept a first generation system.

In almost all evaluation reports in the literatures, we see that MT researchers usually use subjective measures such as "readable", "understandable", "preserve meaning", "preserve meaning and style," and so on, as the measuring criteria for a system. As such, when a system is improved in these aspects, the system is considered improved. However, we found that many end users and customers have another different criterion, namely, "publishable." In other words, since most documents are to be published to the public, high quality translation is required. Furthermore, a customer may ask for a special style of the target document, which is usually different from that of the source document, and prefer special lexicon so that the translated materials, say computer manuals, could be consistent with the style used by the customer for a long time. This means that the customers are more concerned with producing a preferred style for the target language instead of source style preservation, and the customers might have strong demands in adapting the system to their style.

Another important concern to the customers is the translation cost. The cost includes the time to tune the system for the customer's special demands, and the response time to the customer's feedback. If a company is willing to get a system for its own use, the extra cost to pay for tuning the system to its special needs is an important concern. So, if two systems are not significantly different in translation quality, the customer almost definitely prefers to a system that has a larger number of technical terms in the dictionary, better pre-processing and post-editing facilities, and so on. The users rarely want a system simply because it uses an n-th generation technology.

These facts mean that we should have a way to measure system performance in terms of user expectation, instead of a subjective measure for the designers, and try to optimize such a criterion in the target language side. For example, we might have to measure the editing cost in terms of the number of editing operations required to post-edit the raw translation to the final revised version. The designer's view of "readable" or "understandable" is not directly related to customer expectation. Therefore, improving the system performance in terms of the designer's view might not be a direct way to meet customer demands. In

fact, our experiences show that improving the readability of the raw translation is not equivalent to reducing the cost of post-editing; the editing cost is only slightly reduced and sometimes even increased slightly (Su 1992b). The major point is that the stylistic requirements of the customers significantly affect the editing efforts. Therefore, as another important factor, being able to incorporate user feedback is important in system design. Failing to notice user expectation and user feedback might mislead the research directions of the designers, and result in an expensive yet impractical system.

3 What's wrong with the second generation MTs?

We believe that most problems of the second generation MT addressed here may be attributed to three implicit assumptions made in developing a system: (i) a system with improved functional modules, as judged by the designers, would also generate output that gets higher degree of customer satisfaction, (ii) a better architecture, a better knowledge representation or a better grammar formalism would result in better translation quality, and (iii) the cost for constructing a large scale system is only linearly proportional to the size of the system.

The first assumption is questioned in Prof. Somers' criticism on the second generation MT scheme. A stratificational paradigm underlying such designing philosophy tends to preserve source language structures and styles, hence, it may not meet user requirements in target language style directly. As noted in the previous section, we could attribute this problem to the "one-way" designing methodology, where no feedback path from the users or post-editors could be fed to the system, and the system could not be adapted to a customer's flavor from previously edited text. A better way might be to use a bilingual text corpus, probably annotated with structural information, to acquire the corresponding translation parameters bidirectionally from both the source and target sides. Such a "two-way" designing approach (Su 1992b) will enable the feedback from the users to be considered, and make automatic or semiautomatic acquisition of the translation knowledge possible.

The crude comment such as: *"Virtually any theory, no matter how silly, can be the basis of some effective MT"* is a direct questioning against the second assumption. Some MT researchers seem to believe that a better knowledge representation and a better grammatical formalism also imply better translation quality. While it is generally true that a better knowledge representation or grammatical formalism provides a more elegant way to describe linguistic phenomena, it is doubtful how much development cost can be saved in the context of the messy and unrestricted real text. Practically any reasonably designed system could have the formal power of a Turing machine if a procedure driven mechanism is supplemented; even with the old ATN architecture, we don't think it's particularly difficult to implement the same function of the state-of-the-art

grammatical formalism. This means that different systems may have the same expressive power and capacity, and even a first generation system could be trimmed to outperform a second generation system. Therefore, it seems to be more practical to seek for a computational approach that could optimize the system performance in a meaningful sense, rather than to find theoretically interesting alternatives that could not be justified.

As for the third assumption, some researchers might think that a laboratory system that is proved feasible in small scale can be scaled up simply by enlarging the dictionary, knowledge base, and so on; and the cost for constructing a large scale system is only subject to linear increase with the system size. Consequently, the testing results in a laboratory environment often mislead to an over-optimistic estimation of the system performance. The problems resulted from scaling up a system to a large scale system is usually overlooked.

In fact, scaling up a system might subject to the "robustness" problem of an MT system. People familiar with the methods of performance evaluation will appreciate the fact that "performance" is a random variable that needs to be estimated. The estimated performance is affected by database size, model complexity and the performance evaluation method itself. A laboratory system that is over-tuned to a particular set of data might perform satisfactory under the close test environment. However, due to the lack of robustness of its underlying language model, it is very likely to perform badly when put into practical use (Su 1990).

An over-optimistically estimated performance due to an over-tuned knowledge base may also encourage the designers to adopt a complex language model by incorporating whatever information that could be thought of. Consequently, the engineering feasibility to acquire the required find-grained knowledge is usually doubtful for a large scale system. For example, acquiring a fully annotated lexicon of 100,000 entries that consistently supports such a complicated formalism is rarely an easy task. As a result, the cost might not be affordable and we simply get another impractical framework.

4 Why MT research advances so slowly?

In comparison with other communities, the advance of MT research is relatively slow. One important factor that affects the technology advances in MT development is the lack of a common testing environment for the MT community. This means that MT community does not have (i) a set of *objective* and *practical* evaluation criteria for performance evaluation, (ii) a set of *common* databases as the testing suite, and (iii) a *quick* evaluation method.

Because there is no objective and practical criteria for performance evaluation, research groups kept on proposing *alternative* solutions without knowing whether they are *better* solutions. So some ineffective computational models might still be used for a long period of time. Also, impractical measures such

as "readable" and "understandable" based on a designer's view make it hard to optimize a system to meet customer requirements. Under such circumstances, it is not surprised why the second generation MTs are not considered much better.

To make various computational models directly comparable in a more objective way, a set of common testing suite is probably the most important component. People familiar with the speech processing community will appreciate the fact that the series of common tasks setup by DARPA for testing the various speech processing systems have created a competitive environment for the community. It also accelerates the pace of putting the technologies into practical use. The common database forces the research groups to seek for effective computational models and verify their effects under the same testing environment. This stimulates the speech processing community and promotes the advance of speech processing technologies step-by-step from small vocabulary isolated word recognition to large vocabulary continuous speech recognition and spoken language processing. These facts suggest that the MT community (and natural language community) might benefit from the past experiences of the other communities in promoting the technology advances by following the same activities. (Note that some people complained that the common tasks might prevent people from working on more basic research. But we still think it rewardable from a practical point of view.)

Another hindrance factor that block MT development might be the lack of a *quick* evaluation approach. To be objective in performance evaluation, a particular method should be testified against a large testing suite. However, it may take a long time to get the results, for instance, by human inspection. During the development process, performance evaluation might have to be repeated with high frequency to testify various approaches. The lack of a quick evaluation method might either reduce the frequency of testing or prevent the designers from finding wrong research directions earlier. Both will inevitably defer the detection of some ineffective computational models, and block the technology advance in MT development.

5 What should be pursued in future MT research?

The above arguments seem to suggest the following directions to follow. (i) The designer's view should match the user's view in order to improve system performance in the right way. A quality measure that indicates the performance gap between the system capacity and user expectation should be adopted so that the performance measure being optimized reflects real user requirements. It also means to use a "two-way" designing philosophy by incorporating user feedback as a system designing feature, and use bilingual corpora together with appropriate language models to acquire the parameters that govern the transla-

tion mechanism. (ii) The knowledge acquisition problem should be considered more seriously than the knowledge representation problem. This means to develop more effective yet less expensive computational models, such as better scoring mechanisms, for resolving various linguistic problems in various phases; the use of discriminative features, rather than impractically complicated feature structures, should be emphasized in this regard. (iii) Objective and practical evaluation criteria should be defined, a common testing suite should be established and quick evaluation methods should be developed so that ineffective computational models can be detected and discarded as early as possible.

Among the non-orthodoxy approaches in Prof. Somers' list, the corpus-based MT scheme seems to be a good candidate to meet these targets. A corpus-based system could be well parameterized; its performance could be controlled in terms of a set of parameters in a predictable way; effective computational models and objective scoring mechanisms could be established; discriminative features could be found from automatic/semiautomatic training of a large corpus; feedback control can be incorporated with little difficulty; robustness of the system could be improved by using discrimination and robustness oriented adaptive learning approaches (Su 1992a, 92b). Hence, it is much easier to improve such a system under the designer's control. However, purely statistical approaches such as (Brown 1990) are not encouraged because of the lack of high-level language models. A better way might be to apply statistical approaches on top of well justified linguistic models. Such a designing philosophy is called a corpus-based statistics-oriented (CBSO) approach (Su 1992a). A CBSO approach emphasizes the use of well justified linguistic knowledge in constructing the base language model, and uses statistical approaches to acquire the corresponding parameters. As a result, the mechanism does not have to spend precious corpora in training well-know phenomena, and the state space could be kept much smaller. By investigating statistical behavior on top of a high-level language model, the correlation among different constituents could be explored even if they are far away. Therefore, long distance dependency can also be handled much easier. In fact, shifting purely statistical approaches toward CBSO paradigms by using appropriate linguistic models had been observed in some recent reports concerning corpus-based MT.

6 Concluding Remarks

In summary, we consider it most important to establish a common testing suite and some objective evaluation criteria for evaluating various frameworks so that we know what is meant by "much better" before asking "why," and try to optimize the criteria for whatever framework being used. The evaluation criteria should have a unifying view for both the users and the designers so that system improvement is directed toward user requirements. It should also be fast enough so that researchers could find wrong directions as early as possible. Feedback

from users should be seriously considered as a designing feature, and bilingual corpora should be used to acquire the parameters of the underlying translation mechanism, instead of using "one-way" designing approach.

We personally believe that the system designers should pay more attention in integrating the linguistic parts and the computational parts in a way that enable the system to be controlled by a set of parameterized variables. Only when a system scheme is characterized in this way can it be continuously improved under the designer's control, just like any other scientific experiments, even when the system scales up.

7 REFERENCES

Brown, P. F., et al. 1990. "A Statistical Approach to Machine Translation," *Computational Linguistics*, vol 16, no. 2, pp. 79–85. 1990.

Chen, Shu-Chuan, Jing-Shin Chang, Jong-Nae Wang and Keh-Yih Su, 1991. "ArchTran: A Corpus-Based Statistics-Oriented English-Chinese Machine Translation System," *Proceedings of Machine Translation Summit III*, pp. 33–40, Washington, D.C., USA, July 1–4, 1991.

Su, Keh-Yih and Jing-Shin Chang, 1990. "Some Key Issues in Designing MT Systems," *Machine Translation*, vol. 5, no. 4, pp. 265–300, 1990.

Su, Keh-Yih and Jing-Shin Chang, 1992a. "Why Corpus-Based Statistics-Oriented Machine Translation," to appear in *Proceedings of TMI-92*, 4th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, June 25–27, 1992.

Su, Keh-Yih, Ming-Wen Wu and Jing-Shin Chang, 1992b. "A New Quantitative Quality Measure for Machine Translation Systems," to appear in *Proceedings of COLING-92*, 14th Int. Conference on Computational Linguistics, Nantes, France, 20–28 July, 1992.