BehaviorTran: Its Current Status, Prospects and Philosophy

Jing-Shin Chang and Keh-Yih Su

Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 30043, R.O.C.
email:shin@hermes.ee.nthu.edu.tw, kysu@bdc.com.tw

1 Organization and Historical Review

The BehaviorTran (formerly the ArchTran [Su, 90; Chen, 91]) English-Chinese Machine Translation System is the first of its kind research launched in Taiwan, Republic of China; it is also among the first commercialized English-Chinese systems in the world. The first commercialized system was released in 1989. It currently serves as the kernel of a translation service center.

The research on BehaviorTran began as a joint effort between the National Tsing Hua University (NTHU), Taiwan, and the Behavior Tech Computer Corporation (BTC) in May, 1985. It was initiated by Professor Keh-Yih Su of NTHU and was supported by several professional consultants from local universities and research organizations. As the research scaled-up, it was transferred to the Behavior Design Corporation (BTC R&D Center) founded in February 1988 to continue on the improvement of this MT system.

The BehaviorTran MT R&D group is special in that it is not only an MT R&D group, but also featured an in-house co-operative Machine Translation Service Center, which provides commercialized customer services as well as real-task feedback to the MT research team. Currently, about 30 R&D staffs work for the R&D center and about 10 of them belong to the MT research group. The MT translation service center is operated by professional project managers and full-time post-editors. Part time post-editors are linked with the center through public phone network. Such organization provides a good environment to investigate the whole process of real translation tasks under MT environment.

Because of it's strong association with the academic communities, this R&D center is also an active member of the computational linguistics societies in this local area and in several international associations. More than 50 technical papers in international conferences, professional MT journals and general NLP journals had been published. The publication of technical papers is a regular activity here. It is also one of the founding members of the ROCLING Society (the Computational Linguistics Society of the Republic of China) as well as an active member in the related activities. In addition, there are several international cooperations or resource sharing contracts with a few famous research Labs in the North America area.

Instead of providing stand-alone machine translation systems, the target is to provide users an environment very much like an in-house translation environment. This policy is adopted ever since its foundation because successful commercialized MT operations in the world suggest that in-house translation is probably the best way for providing translation services to customers who really want a "solution" without the overhead for maintaining an MT system by themselves.

As a result, a VAN-based translation service center is set up. In such a service configuration, electronic articles are transmitted to or from the customers, the service center, and the post-editors *via* the local or public data network. This greatly reduces the overhead of the customers in terms of time, cost, and security assurance. The clients are also relieved of the overhead of maintaining the knowledge bases (e.g. dictionaries) by themselves.

Currently, BehaviorTran has established a customer base including several internationally renowned computer companies. Most of these customers are among the top-rank and well-known software, hardware, PC or workstation companies in the world.

The primary domain for BehaviorTran is computer manuals and related documents; other technical fields, like mechanical fields and medical fields, are also supported to a wide variety of private and government organizations.

The system is developed on a UNIX workstation environment; it was written in the C language. The raw translation is mostly post-edited on PCs, which are connected to the translation server through the network of translation service center.

Sophisticated support tools are also packaged into the translation workstations, integrating aids such as OCR,

special-purpose text editor, in-house glossaries, bi-texts, DTP, and so on. Such supporting softwares are provided in part or in full by the other R&D groups of the BDC. These facilities include a Chinese-English bilingual writer's workbench, which is characterized by a full-featured, Chinese-English bilingual desktop publishing system for text, equation, table, graphics and image.

2 General Technical Perspective

The research of BehaviorTran started with a conventional transfer-based MT architecture. Many rules are encoded in the system to take care of the various linguistic problems. However, as the system scaled up, it is found that such a rule-based approach would suffer from some problems. In particular, it lacks an objective preference measure to deal with uncertainty knowledge. In addition, it is hard to deal with complex and irregular knowledge. Exceptions to the rules occur from time to time. In the system engineering level, it is hard to maintain the consistency of the large amount of fine-grained knowledge among different persons at different times. There is no systematic way to acquire linguistic knowledge as proposed in various literatures. The acquisition of the large amount of fine-grained knowledge with human intervention is thus costly and time-consuming.

Therefore, the focus is soon moved to corpus-based approach after 2 years' research works [Su, 92]. The main design features of the BehaviorTran system is therefore an integration of the conventional MT methods and the new corpus-based technologies, which is now known as the "hybrid method" in the MT communities. Under such an architecture, the BehaviorTran is featured with the adoption of a mixed parsing strategy (called *bottom-up parsing with top-down filtering*), a scored parsing mechanism, and the *corpus-based statistics-oriented* (CBSO) paradigm for linguistic knowledge acquisition.

Since the research of the BehaviorTran system is characterized as being corpus-based and statistics-oriented, research directions are toward designing systematic and automatic methods for acquiring language model parameters, and toward using preference measure with a uniform probabilistic score function for ambiguity resolution.

In developing CBSO language models, special emphases are placed on using discriminative features, in stead of full-blown semantic analysis, for ambiguity resolution. This philosophy leads to useful corpus-based technologies which possess the required ambiguity resolution capability at a much smaller cost in knowledge acquisition and maintenance. In the computational parts, the CBSO approaches of the BehaviorTran place emphases on robustness of the computational models so that the language parameters can be applied to unseen text as well.

With all these in mind, the BehaviorTran R&D team had established the various preference measures in lexical analysis, syntactic analysis, semantic analysis, probabilistics transfer model and generation model, all under a unified probabilistic framework. Discriminative and robust oriented techniques are also established to make better estimation on the various language parameters for these models.

3 Toward Bidirectional Transfer with Bilingual Corpora

Since the operation of the BehaviorTran, we found that most customers have a strong demand for the translation output to be "publishable" to the public. Furthermore, a customer may ask for a preferred style of the target document, which is usually different from that of the source document, so that the translated materials could be consistent with the style long developed by the customer. 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 styles. The time to tune the system to fit the customer's special demands, and the response time to the customer's feedback are also very important concerns to a customer.

One major reason why present MT systems are not yet widely used lies in the fact that the generated output is strongly affected by the source language; the style simply does not follow what a native speaker of the target language would expect [Su, 93].

This is not surprising since conventional transfer and generation processes of an MT system depend heavily on hand-coded transfer rules on particular transfer units. The transfer operations are essentially uni-directional in that transfer operations are designed from the source language to the target language; the transfer rules are not trained to acquire the best mapping between syntactic structures that are produced by the source grammar and a "legal" and "natural" target grammar.

There is also the problem for identifying the appropriate transfer units and the canonical sequence of the transfer operations with such approaches. Without a systematic approach for localizing the transfer units to a finite set of primitive units, it is hard to establish the transfer rules or transfer operations completely and effectively. Furthermore, since most of such processes cannot be parameterized, a customer's feedback could not be used as input for fitting the customer's particular style. It is thus hard to tune such systems to a customer's preference by providing a preference score to each possible transfer operation.

The goals of the BehaviorTran team are thus to construct a *bidirectional transfer and generation model* [Chang, 93] between the two languages, and make the transfer and generation process a highly parameterized and feedback controlled system with automatic approaches for acquiring the transfer units and operations.

To achieve such goals, our method is to reduce the annotated syntax trees of a translation pair, which are produced by a source grammar and a target grammar, respectively, to a *normalized* version of these annotated syntax trees. The primitive transfer units between such tree pairs are then acquired with an automatic approach. A transfer score and a generation score are then defined in terms of such localized transfer units to find the best or preferred mapping.

With such a model, the transfer operations can be limited to only a finite set of transfer units. Furthermore, since the normalized syntax trees are produced by their respective grammars, the target structures selected by this mechanism will not be influenced by the source grammar. Therefore, the mapping is essentially bidirectional, which could be started from either side.

By decomposing the syntactic structures into primitive transfer units and using a uniform probabilistic model for the transfer and generation scores, it is easy to parameterize the transfer and generation process. With such a parameterized transfer model, the BehaviorTran's approach further has the potential capability of tuning the transfer patterns to the style of a particular customer. We therefore are strongly in favor of such an approach.

4 Future Goals

After almost one decade of research and development in machine translation, and through the commercialized operation of the BehaviorTran MT system, we believe that new corpus-based approaches will play a very important role to make machine translation more practical and widely used.

In sum, we think that humans are competent in general language modeling while computers are effective in processing massive data. Therefore, it is appropriate to take advantages of well-recognized linguistic phenomena, setup probabilistic language models by humans, and estimate the language parameters of the probabilistic models from large corpora with well-established techniques from the statistics communities.

Under this direction, a few advanced technologies are now under development. In particular, the research works will be directed toward automatic corpus conversion, annotation, automatic model refinements and better parameterized language models.

References

[Chang, 93] Chang, J.-S. and K.-Y. Su, "A Corpus-Based Statistics-Oriented Transfer and Generation Model for Machine Translation," *Proceedings of TMI-93*, pp. 3–14, 5th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, July 14–16, 1993.

[Chen, 91] Chen, S.-C., J.-S. Chang, J.-N. Wang and K.-Y. Su, "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, 90] Su, K.-Y. and J.-S. Chang, "Some Key Issues in Designing MT Systems," *Machine Translation*, vol. 5, no. 4, pp. 265-300, 1990.

[Su, 92] Su, K.-Y and J.-S. Chang, "Why Corpus-Based Statistics-Oriented Machine Translation," *Proceedings of TMI-92*, pp. 249–262, 4th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, June 25–27, 1992.

[Su, 93] Su, K.-Y. and J.-S. Chang, "Why MT Systems Are Still Not Widely Used?" *Machine Translation*, vol. 7, no. 4, pp. 285–291, Kluwer Academic Publishers, 1993.

Published in *Journal of the Asia-Pacific Association for Machine Translation*, **No. 7, 1994.**