Corpus-Based Statistics-Oriented (CBSO) Machine Translation Researches in Taiwan

[+]Jing-Shin Chang and [*+]Keh-Yih Su

[*]Behavior Design Corporation
2F, No. 5, Industrial East Road IV, Science-Based Industrial Park
Hsinchu, Taiwan 30077, R.O.C.

[+]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

Abstract

A brief introduction to the MT research projects in Taiwan is given in this paper. Special attention is given to the more and more popular corpus-based statistics-oriented (CBSO) approaches in MT researches. In particular, the parameterized two-way training philosophy in designing the second generation BehaviorTran, which is the first and the largest operational system in this area, is introduced in this paper.

1. Overview of the Machine Translation Projects in Taiwan

The first MT research project (the ArchTran MT system, aka the BehaviorTran system) in Taiwan was conducted in 1985 in National Tsing-Hua University [Hsu 86]. Later, MT related projects were subsequently started in many other academic institutions and private sectors, including the Institute of Computer Science and Information Engineering of National Taiwan University, Industry Technology Research Institute (ITRI), Matsushita Electric Institute of Technology (Taiwan), Wang Corporation (Taiwan), Institute of Information Engineering of National Chiao-Tung University, Institute of Information Science, National Tsing-Hua University, and so on. Related NLP researches are also conducted in other Institutions, such as the Institute of Information Science, Academia Sinica and the Telecommunication Technology Laboratory.

Because of the blooming growth in the related researches, an annual conference, the ROCLING conference, was initiated in 1988; the ROCLING Society (ROC Computational Linguistics Society) was also founded as a regular organization for promoting various NLP researches, including machine translation. Many workshops were called in a non-regular basis since the foundation of the ROCLING Society. SIGMT for machine translation, and special MT workshops, in particular, were also held for people who are interested in this technical field of research.

Since the BehaviorTran MT system is the first MT project in this area, and it is also the largest operational system in this area, its working experiences have significant influence on the other related researches. In fact, the adoption of corpus-based statistics-oriented approaches for developing its second generation system for the last few years goes roughly in parallel with the corpus-based research trend in the worldwide computational linguistics community; such trend is also appreciated by the other research institutions in this area. Therefore, we will follow the development track of the BehaviorTran system and outline its current designing philosophy for the second generation BehaviorTran in this paper. In particular, we will introduce its two-way training philosophy for automatically acquiring the required translation knowledge with bilingual corpora in the following sections.

1.1. The BehaviorTran MT System

The BehaviorTran MT system, the first MT project conducted in Taiwan, was founded by the Behavior Tech Computer Corporation (BTC) as a joint project with the Department of Electrical Engineering, National Tsing-Hua University (NTHU). The project started from May 1985 under the supervision of Professor Keh-Yih Su, National Tsing-Hua University, and was transfered to the Behavior Design Corporation (BDC) at the Science-Based Industrial Park, Hsinchu, Taiwan in February 1988. The translation service center was established in 1989, which serves to provide in-house translation services for many international companies in the areas of computer software, cars, mechanical industry, and so on.

Instead of providing stand-alone machine translation system, the target of BehaviorTran is to provide 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. Currently, the primary domain for BehaviorTran is computer manuals and related documents; other technical fields, like mechanical fields and chemical fields are also supported to serve a wide variety of private and government organizations.

The BehaviorTran MT system has a transfer-based architecture, which was constructed on top of an Augmented Phrase Structure Grammar and an extended LR parser (capable of bottom-up parsing, top-down filtering, and partial parsing for short or incomplete sentences.) In addition to rejecting unlikely analyses with lexical, syntactic and semantic constraints, the system is also featured with statistical modules for scoring various analyses in order to get the best possible analysis in the analysis phase.

The dictionaries of the BehaviorTran are stratified into general dictionary, idiom dictionary, split idiom dictionary (for idiomatic expressions or collocates that are not groups of adjacent words), technical domain dictionary, user dictionary, and project dictionary. The dictionaries are unified into one before any translation project begins, with the project dictionary taking the first priority, and the general dictionary taking the lowest priority (following the reverse of the above-mentioned order.)

Since 1987, many of the system modules are enhanced with statistical scoring mechanisms, and the large amount of fine-grained linguistic knowledge is acquired with corpus-based statistics-oriented methods. More details on its current development trend and philosophy are given in the following sections.

2. Technical Bottlenecks and Current Strategies

It was well-known, during the past 50 years, that fully automatic high quality machine translation is quite difficult to develop. The major problems are attributed to the large amount of fine-grained linguistic knowledge required for high quality translation, and the constantly increasing vocabularies in the real world. The translation quality is also significantly influenced by the traditional one-way source-dependent designing approaches in acquiring the translation knowledge [Su 93].

The research of the BehaviorTran started with a conventional transfer-based MT architecture. Many rules are encoded in the system to take care of the various linguistic problems. As the research extends to a large scale system, it is found that such a rule-based approach would suffer from many knowledge acquisition problems, which gradually become the most important technical bottlenecks to be defeated in the second generation BehaviorTran.

The most serious problem, from a user's point of view, is that the target translation is usually too literal and its style is usually bounded by the source language [Su 93]. To have a customer-satisfied system, it is therefore desirable to adopt a model to prevent the translation from being too literal. Besides, since knowledge acquisition and domain adaptation are also the major bottlenecks in real commercialized machine translation systems. A parameterized approach is thus adopted in developing the second generation BehaviorTran for attacking such problems, so as to enable a machine translation system with high modularity and to acquire its translation knowledge from a bilingual corpus with a two-way training method.

2.1. Difficulties of one-way design systems

In most one-way design systems that follow an analysis-transfer-generation process, the major problem is that the generated sentences are often strongly bounded to the analysis grammar of the source language since the generation grammar is often a slightly modified version of the source analysis grammar, which is influenced greatly by the source language. The generation grammar might preserve lots of stylistic characteristics of the source language such that the generated sentences are unnatural to the native speakers. As a result, the translated sentences will be too literal for a native speaker. Because almost all the translation knowledge in a one-way design system is derived, explicitly or implicitly, based on the training corpus of the source language, it is difficult to learn proper transfer knowledge that governs the translation of the source sentences into the most preferred target sentences.

Furthermore, a source-dependent system is hard to make the system reversable. Therefore, many target language informations and modules may not be reused when it becomes the source language and *vice versa*. And this drawback becomes more and more salient since BehaviorTran intends to extend itself to a multilingual translation system.

2.2. Difficulties of knowledge acquisition

Under traditional rule-based MT architecture, the linguistic knowledge is acquired by induction from various observations, the acquisition of the large amount of fine-grained knowledge with human intervention is thus costly and time-consuming. To relieve the burden in knowledge acquisition, symbolic learning methods had been tried in the literatures to organize the linguistic knowledge. However, such approaches are usually awkward in handling uncertain knowledge and do not gain much success so far.

From a system engineer's point of view, traditional rule-based approach is hard to maintain the consistency of the large amount of fine-grained knowledge among different persons at different time. Since there are no objective preference measure to deal with complex and irregular knowledge, exceptions to the rules occur from time to time. Therefore, it is desirable to use a stochastic model for reducing the labor in knowledge acquisition.

2.3. Difficulties of domain adaptation

As the translation domains of BehaviorTran extend from computer science to many other domains, such as electrical engineering, mechanical engineering, aviation and navigation, and the number of customers and posteditors are increasing, it was found that operating an MT system in different domains is not simply a task of changing the domain-specific lexicon. The difference in syntax and semantics among different domains usually make a general translation system hard to generate high quality translations steadily, even if different sets of lexicon have been attached during the transfer process. To render good translations steadily, not only the lexicon of different domains, but also the analysis rules, disambiguation rules, transfer rules and generation rules should be modified according to the changes in various domains. However, it will be costly to re-acquire the translation knowledge for each translation domain. Thus, how to develop the techniques for effectively adapting the system for different domains without re-acquiring the transfer knowledge from scratch is an important issue to be addressed in the second generation BehaviorTran.

3. New approaches in the second generation BehaviorTran

The bottlenecks in developing a large MT system is mainly the acquisition of the underlying translation knowledge (including likelihood probabilities for non-deterministic or uncertain linguistic knowledge) and the adaptation to different domains. To attack the drawbacks of traditional MT system, the design philosophy of the new generation BehaviorTran moves toward a parameterized system with a two-way training method. The main concern for *two-way* training is to avoid generating source-dependent output, and the main concern for using a highly *parameterized* system is to make knowledge acquisition and domain adaptation relatively easy.

3.1. Architecture in the second generation BehaviorTran

A schematic view of the translation flow in the new generation BehaviorTran is shown in Figure 1 below. In this figure, S and T represent the source and target sentences respectively. PT stands for the parse tree, NF1 stands for the first-level normal form, and NF2 stands for the second-level normal form. And the subscripts 's' and 't' attached to the above symbols stand for the source and target language respectively. The circles on the source side, denoted by Gs, NR1s, NR2s represent the source grammar (Gs), normalization rules (NR1s, NR2s), which serve to normalize the parse tree into the first-level and second-level normal forms. The circles on the target side, on the other hand, represent the generation rules (GR2, GR1, GR0) of various levels which are the counterparts of the normalization rules on the source side. In addition, P(X|Y) represents the conditional probability for X to appear given that Y is observed. Such parameters (conditional probabilities) are used to assign preference scores for disambiguation.

In this flow, there are several intermediate representations, referred to as the Normal Forms (NF), which are normalized constructs from the parse trees. Figure 2 and 3 are examples of such normalized constructs. The NF1 tree (normalized syntax tree), for instance, is acquired by normalizing various syntactic variants into the same form; and the NF2 tree is a kind of normalized structure obtained by normalizing certain semantic variants into the same construct, by applying certain normalization rules (NR2). With the introduction of the NFs, we intend to set up a set of linguistically-justified intermediate levels which can divide the original translation process into several independent phases.

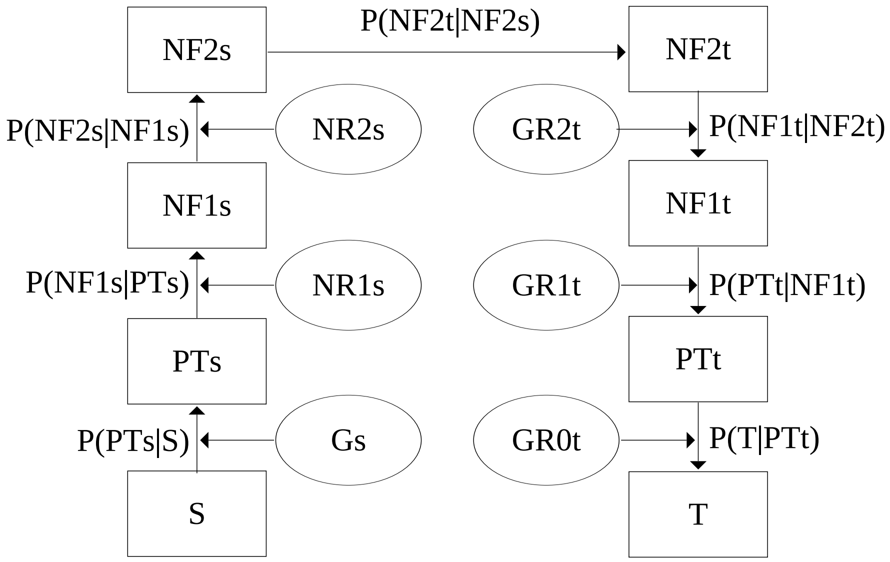

Figure 1 A schematic view of the translation flow in the second generation BehaviorTran

```
SIMP -------------------------------------------------------
  |                              \   \                      \
  SB -----------                  V   NP ---                SB ---------------
  |      \       \                |   |    \                 |        \        \
  |      V       NP ------        |  DET   NP ------         |        V        VP ----------
  |      |       |        \       |   |     |        \       |        |        |    \        \
  |      |       |         |      |   |     |        |       |        |        |    NLM    NP
  |      |       |         |      |   |     |        |       |        |        |     |        |
 comp    v       n         n      v  art    n         n     cjsb      v        v    quan      n
  To  meet spectrum  specific  allow  a 30-minute warm-up  before attempting  make  any calibrated
           -analyzer  -ations                                                        -measurements
```

Figure 2 Example: an NF1 Tree (Normalized Syntax Tree)

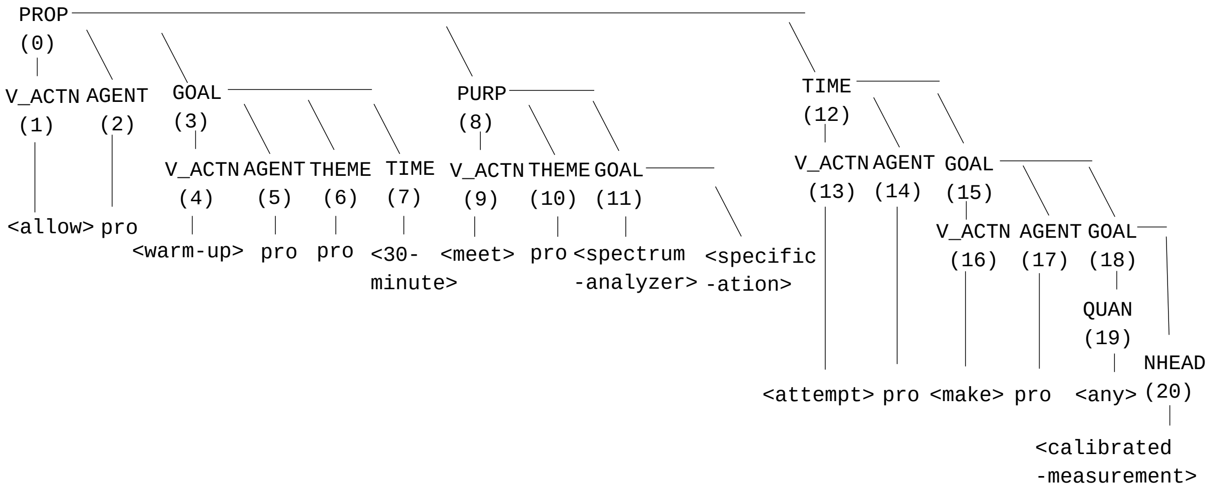

Figure 3 Example: an NF2 Tree (Normalized Semantic Tree)

Note that in such a system, the phrase structure grammar, the normalization rules and the generation rules only produce possible parses or normalized constructs without involving in the disambiguation tasks; the system parameters (i.e., the conditional probabilities), on the other hand, play the major role for disambiguation or selection of preferred constructs based on quantitative preference scores [Chang 93, Su 95].

Since the disambiguation process and the required parameter values are acquired from large training corpora, it will be relatively easy to acquire and maintain without much human cost. In the following sections, we will indicate in more detail the advantages of the second generation BehaviorTran which adopts such a new architecture with parameterized MT approaches and a two-way training method.

3.2. Parameterized BehaviorTran

In contrast to the conventional systems, a parameterized system, as described in the last section, is characterized by a quantitative optimization criterion and a training mechanism for acquiring the language parameters (such as a set of probablities or scores) from real text corpora. The training mechanism is simply an estimation process to get the parameter sets from a corpora according to some objective optimization criteria. And such methods are the essential principle of the parameterized approaches [Hsu 95, Su 95]. In contrast to most other systems whose knowledge is transformed from existing linguistic theory, a parameterized system as characterized here could have many potential advantages as described below.

First, parameter learning is usually easier and more objectively optimized than symbolic learning approaches. A parameter learning process usually involves only mathematical computation instead of complicated induction mechanisms. Therefore, the driving mechanism is simple and each learning step could be quantitatively controlled. Furthermore, the search path toward the best parameters for most optimization criteria could be implemented easily by adjusting the parameters iteratively according to incorrectly analyzed sentences [Amari 67].

Second, the parameter sets could be easily adjusted for the various styles in different domains in a systematic way. The knowledge acquisition cost, in terms of man-years, is usually smaller. A parameterized system thus provides the potential benefits of using alternative sets of parameters for different applications, and leaves the driving mechanism, functional modules (or the knowledge base) the same. Therefore, a parameterized system is preferred in terms of knowledge acquisition cost and adaptability to different requirements [Su 95].

From the description above, it clearly shows that parameterized approaches will allow both the transfer knowledge and the generation knowledge to be acquired systematically from the corpus [Chang 93, Su 95]. With this approach, linguists will be requested to construct the language model instead of maintaining a large set of rules by hands; corpora are used as the main information source and statistical techniques are used to learn model parameters and the system will automatically acquire the knowledge from the corpus. Thus, uncertainty or preference can be interpreted objectively and consistently and the burden of rule induction will be moved from linguists to machine.

In addition, with such a model, the transfer operations can be limited to only a finite set of transfer units [Chang 93]. And by decomposing the syntactic structures into those primitive transfer units and using a uniform probabilistic model for the transfer and generation scores, the most preferred rules and the parameters of the underlying language model can be easily adapted to different domains. Thus with such a parameterized transfer model, the BehaviorTran's approach will have the potential capability of tuning the transfer patterns to respective styles or domains for a particular customer. We therefore are strongly in favor of such an approach.

3.3. Two-way training

As mentioned previously, one major problem with conventional MT systems is that the target translation depends heavily on the analysis grammar of the source language. Although the architecture in the translation flow in Figure 1 provides a good framework for a parameterized MT system, it does not guarantee to remove such source dependency if the translation knowledge is not acquired in a proper way. In particular, if the translation knowledge is acquired following the conventional analysis, transfer and generation flow, it will still be a one-way system, whose translation knowledge for generating the target sentences will still be influenced significantly by the source language.

To change the system architecture from one-way design toward two-way design, the transfer knowledge should be trained from both properly normalized source and target knowledge representations, which should both fall within the range of the sentence that will be produced by the native post-editors, according to the respective discourse context of the source language and target language, respectively. The following flow shows the general ideas for training a two-way system. The bold arrows at the right hand side emphasize that the intermediate representations for the target language are directly derived from the target sentences in an aligned bilingual corpus, according to the target grammar (Gt) and target normalization rules (NR1t, NR2t) (which are simply the reverse of the generation rules GR0t, GR1t and GR2t in Figure 1.)

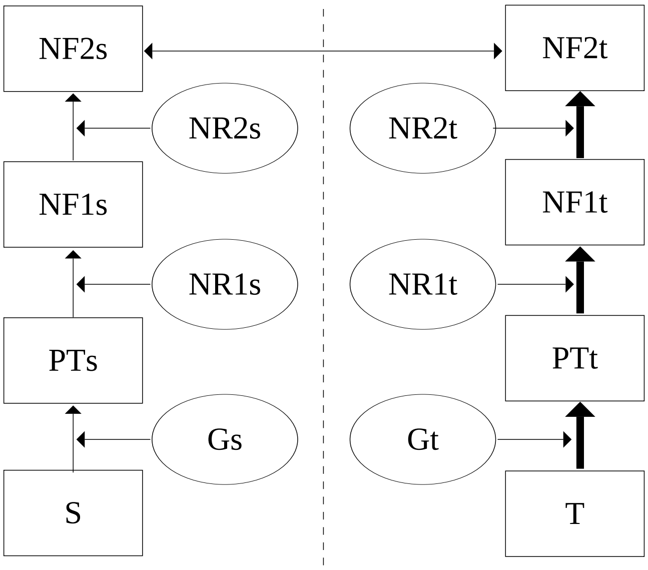

Figure 4 Training the translation knowledge in a parameterized two-way training system.

The arrow symbols in Figure 4 indicate that the PT's, NF1's and NF2's for both the source and target sentences are derived from the source and target sentences respectively, based on their own phrase structure grammar and normalization rules. Therefore, all such intermediate representations are guaranteed to fall within the range of the sentences that will be produced by the native speaker of the source and target language; the transfer components only select those preferred candidates among such constructs. Once the normalization and transfer knowledge is learned based on the normal forms of the source and target languages respectively, then, high quality translations will be generated by such an approach, since the normal forms will no more depend on the grammar of the other language.

As mentioned above, one of the reasons why most MT systems are not widely used today lies in the fact that the generated output is strongly affected by the source language, and the style simply does not follow what a native speaker of the target language would expect. But with such a two-way training model, this problem could be solved to a large extent. It is expected that the most preferred translations between a language pair can be selected and tuned to follow the grammar and style of the target language using such a two-way training approach, because the mapping is essentially bidirectional, which could be started from either side, and the normalized syntax trees are produced by their respective grammar to prevent the target structures from bounded by the source grammar.

4. Conclusions

In this paper, we present the current status and the future prospects of BehaviorTran, and also introduce the parameterized two-way design philosophy and architecture in the new generation BehaviorTran. With a parameterized paradigm, many traditional transfer-based MT modules can be parameterized to gain more flexibility and better performance, in terms of knowledge acquisition and domain adaptability. The training method for estimating the parameters of the system, based on a two-way design philosophy, is also exploited to acquire the underlying translation and transfer from a bilingual corpus so that source dependency in traditional one-way design could be relieved. And with the superiority in knowledge acquisition, domain adaptation, and source independence, we believe the new system implemented under the new architecture will play an important role in solving the problems encountered in traditional MT systems.

Reference

[Amari 67] Amari, S. "A theory of adaptive pattern classifiers", *IEEE Trans. on Electronic Computers*, vol. EC-16, no.3, pp. 299-307, June 1967.

[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.

[Hsu 86] Hsu, H.-H., and K.-Y., Su, "A Bottom-Up Parser in the Machine Translation System with the Essence of ATN," *Proceedings of International Computer Symposium (ICS)* 1986, Vol.1 of 3, pp 166-173, Tainan, Taiwan, R.O.C., Dec 17-19, 1986.

[Hsu 95] Hsu,Y.-L. Una, and K.-Y., Su, "The New Generation BehaviorTran: Design Philosophy and System Architecture," *Proceedings of ROCLING VIII*, pp. 65-79, Chongli, R.O.C. August 18-19, 1995.

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

[Su 95] Su, K.-Y., J.-S. Chang, and Y.-L. Una Hsu, "A Corpus-based Two-way Design for Parameterized MT Systems: Rationale, Architecture and Training Issues," *Proceedings of TMI-95*, pp. 334-353, 6th Int. Conf. on Theoretical and Methodological Issues in Machine Translation, Leuven, Belgium, July 5-7, 1995.