Two Types of Adaptive MT Environments

Sergei NIRENBURG
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA, USA

Robert FREDERKING
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA, USA

David FARWELL
Computing Research Lab
New mexico State University
Las Cruces, NM, USA

Yorick WILKS
Department of Computer Science
University of Sheffield
UK

ABSTRACT

A number of proposals have come up in recent years for hybridization of MT. Current MT projects — both "pure" and hybrid, both predominantly technology-oriented and scientific (including those currently funded by NSF) are single-engine projects, capable of one particular type of source text analysis, one particular method of finding target language correspondences for source language elements and one prescribed method of generating the target language text. While such projects can be quite useful, we believe that it is time to make the next step in the design of machine translation systems and to move toward adaptive, multiple-engine systems. We describe the architecture of an adaptive multi-engine MT system which uses each of the engines under the circumstances which are most favorable for its success.

1. Multi-Engine MT Architecture

Recent years have witnessed a shift in the balance of scientific and technological efforts in the area of machine translation. All the latest methodological novelties in this field are essentially technology-oriented and do not aim at advancing our knowledge about either basic mechanisms of text comprehension and production or computer models simulating such mechanisms.

The two most recently popular technological paradigms in machine translation — example-based translation (EBMT) and statistics-based translation (SBMT) — require linguistic knowledge only as an afterthought. While the representatives of the above paradigms are still at the stage of either building toy systems (e.g., Furuse and Iida, 1992; McLean, 1992, Jones, 1992, Maruyama and Watanabe, 1992) or struggling with the natural constraints of approaches that eschew the study of language as such (e.g., Brown et al., 1990), a number of proposals have come up for some hybridization of MT. In some such approaches, corpus analysis is used for tuning analysis and transfer grammars (e.g., Su and Chang, 1992). In others, a standard transfer-based approach (TBMT) is followed using traditional analysis and generation techniques but having a transfer component based on aligned bilingual corpora (Grishman and Kosaka, 1992). Still others, use statistical information as the source of preference assignment during text disambiguation (e.g., the outline presented in Lehmann and Ott, 1992). Statistical modeling can be used at some stages of a knowledge-based MT (KBMT) system (see, e.g., Helmreich, 1994).

Current MT projects — both "pure" and hybrid, both predominantly technology-oriented and scientific are single-engine projects, capable of one particular type of source text analysis, one particular method of finding target language correspondences for source language elements and one prescribed method of generating the target language text. While such projects can be quite useful, we believe that it is time to make the next step in the design of machine translation systems and to move toward adaptive, multiple-engine systems.

Practical MT systems are typically developed for a particular text type (e.g., weather reports, financial news articles, scientific abstracts) and for a particular end use — e.g., assimilation or dissemination of information. Special cases, such as translating an updated version of a previously translated text, abound in the real-world practice. Gains in output quality and efficiency can be expected if a machine translation environment can be made to adapt to a task profile. Thus, for example, for translating abstracts of scientific articles in order to select just those that are of particular interest to a customer, a statistics-based approach might be most appropriate. Example-based translation seems to be most promising for translating new versions of previously translated documents. This correspondence between technique, input text type and end use (or output text type) provides further motivation for moving toward adaptive, multiple-engine systems.

We perceive two approaches to adaptivity in MT. Both presuppose an MT environment in which a number of MT engines are present — for instance, one (or more!) each of KBMT, SBMT, EBMT and TBMT engines can be used. In one of the approaches all available engines are "unleashed" on an input text and the final output is assembled from the best text segments, irrespective of which engine produced them. We call this approach the Best Output Segment (BOS) approach. In another approach a heuristic "dispatcher" decides which of the available engines holds the highest promise for a given input text and then assigns the job to that engine. This is the Dispatcher-Based (DB) approach. The BOS approach involves more processing but allows an *a posteriori* selection of the best results. The DB approach saves cycles but relies on heuristic *a priori* selection of the best output. In this latter case, the quality of the dispatcher module is crucial, but additionally, the DB approach expects each of the component engines to be of rather high quality, since they would not (as is the case in the BOS approach) be "bailed out" by other engines in case of failure.

In what follows we briefly describe our first experiment with the BOS approach and discuss the requirements for the DB approach.

2. The Best Output Segment Approach to Adaptivity

Our BOS approach experiment was carried out for a Spanish – English translation set-up in the framework of the Pangloss MT project (Pangloss, 1994) and used three MT engines — KBMT, EBMT, and TBMT.

The KBMT engine we used was the mainline engine of the Pangloss system, a traditional KBMT environment described in some detail in (Pangloss, 1994). It was important for the BOS experiment that this engine generated an internal quality rating for each output segment it produced.

The basic idea of EBMT is simple (cf. Nagao, 1984): an input passage S is compared with the source-language "side" of a bilingual text archive, where text passages are stored with their translations into a target language (or a set of such). The "closest" match, passage S' is selected and the translation of this closest match, the passage T' is accepted as the translation of S. Our EBMT engine used a 100MB bilingual Spanish - English archive of UN official documents. In preparation for processing, the archive was aligned at the sentence level. The matching of input passages with the Spanish side of the archive was allowed to be inexact. Penalties were assessed for omitted and extra words, word occurrences in different morphological forms and differences in word order. The English string translating the best Spanish archive candidate was then found in the English sentence aligned with the Spanish sentence in which the best match candidate appeared. A Spanish - English MRD was used in determining translations of individual words inside the candidate segments. A special routine then calculated the expected quality of the resulting translation, which helped at the result integration stage of multi-engine MT system operation. Our EBMT approach is described in Nirenburg et al., 1993 and Nirenburg et al., *submitted*).

Our transfer system was very simple. It was based on direct lexical substitution fo English words and phrases for Spanish words and phrase, fortified with morphological analysis and synthesis modules. The process relied on a number of databases – a Spanish - English MRD, the lexicons used by the KBMT engine, a large set of user-generated bilingual glossaries as well as a gazetteer and a list of proper and organization names. The user-generated glossaries for our experiment contained about 174,000 entries. Glossary entires contained variables to allow feature matching and indices to link the parts of phrasal entries that translated one another. For instance, the following glossary entry

```
absolver<1> a <dop:2> de
                <poss:2> promesa
==>
release<1> <dop:2> from
                <poss:2> promise
```

can help to generate such English sentences as

```
I release you from your promise;
He released me from my promise;
You will be releasing her from
her promise;
etc.
```

In the rule above *dop* stands for "direct object pronoun" and *poss* for "possessive." Tables of feature correspondences were prepared to make the translation possible. Note that in many cases Spanish features and English features were quite different (notably, for verbs). The numbers in angular brackets are indices which show the morphological synthesizer which word to put in a particular form at generation time. In this experiment we used variables for the following word classes: proper names, such as individual, company and place names; numbers and the various classes of pronouns — personal, possessive, reflexive, direct object, indirect object and possessive absolute.

2.1. Combining Results

The crux of the BOS method is combining results from individual engines. A chart data structure was used to combine results from the individual engines. Before the translation process, the edges of the chart were made to correspond to individual words in the input. New edges are added to the chart through the operation of the three MT engines labeled with the translation of a segment of the input string and indexed by this segment's beginning and end positions. The KBMT and EBMT engines also carried a quality score for each output element.

After all the engines finished their work it is necessary to find the sequence of translation candidates which a) cover the input string as densely as possible (so that there is a translation for as many source text elements as possible); b) use the "best" of the available candidates.

To find the best candidates three heuristics were used – a) internal quality ratings produced by the KBMT and EBMT engines; b) static relative quality assessment of the particular engines we used and c) the length of the translation segment (the longer, the better). Enhancing the quality of these heuristics and generally finding more sophisticated ways of combining findings of individual engines is the most important direction of improvement of our BOS system.

The *chart walk* algorithm producing the final result of the BOS system used the above heuristics. The algorithm uses dynamic programming to find the optimal cover (a cover with the best cumulative score), assuming correct component quality scores. It is described in some detail and illustrated in Nirenburg and Frederking, 1994 and Frederking and Nirenburg, *submitted*.

3. The Dispatcher-Based Approach to Adaptivity

In this approach, a dispatcher module is used to break up the input text into segments and assign each segment to one or another of the available MT engines. Among the possible diagnostics for the dispatcher are:

- Type of translation — whether the result of translation is intended for dissemination or for assimilation; whether a complete translation is needed or an abstract or even a simple categorization of a text (e.g., as a text that is important enough to be translated in its entirety).
- Availability of parallel text in a particular domain and on a particular topic. This is the crucial enabling condition for EBMT and SBMT.
- Amount of ambiguity in the source passage, both in the source language itself and vis-a-vis a target language. The smaller the degree of ambiguity, the more attractive the KBMT approach.
- Size and quality of available KBMT resources (ontology, lexicons, etc.).

The work on the dispatcher, thus, includes a) evaluating the translation context with respect to the four criteria above and b) putting together a decision mechanism which will establish the relative appropriateness of each of the available engines for treating an input passage in a given context. An additional important parameter in the operation of the dispatcher is determining the most appropriate size of input passage to be dispatched to an MT engine. Since an entire input text can be processed by a combination of MT engines, it is necessary to maximize the expected quality of output over a variety of possible ways of "chunking" the input text for processing. This has some similarity with the chart walk in the BOS approach.

The dispatcher will use an additional set of diagnostics determined by the structure of the specific MT engine. The development of these dispatcher heuristics — in other words, how the dispatcher is to be trained (see below) — is a key point of the proposed research. A preliminary analysis of these specific diagnostic heuristics, ordered by the particular engine, follows.

An additional diagnostic heuristic for SBMT inspects the frequency of occurrence of each individual input string item in the corpus. The greater the frequency of the items contained in the text, the greater the likelihood that the SBMT engine will produce good quality output.

The above heuristic will also serve the EBMT engine. A heuristic useful specifically for EBMT is the amount of overlap of an input text with a document already in the source language side of the bilingual archive.

The diagnostics for the TBMT and KBMT approaches mostly check the coverages of appropriate static knowledge sources — grammars and lexicons.

The diagnostics proposed above vary in cost, both in terms of developing the procedures and in terms of their computational complexity. Relatively inexpensive are diagnostics based on recognizing individual forms or patterns in the input (e.g., checking the availability of items in a lexicon or a corpus, checking the length of segments, checking for local sequencing patterns of forms). Somewhat more expensive are diagnostics based on assignment of categories to forms. It is serendipitous, however, that the more costly diagnostics are generally related to initial stages of processing necessary in most engines. This opens a potential for interleaving the processing by individual engines with the operation of the dispatcher.

4. Future Work

The questions of how to optimize the combination of evidence in the BOS approach and how to train the dispatcher in the DB approach are very close to a key question in modern MT: how an MT system is to be evaluated (even as a small-scale proof of concept). We plan an experimental study to improve the procedure for the combination of evidence from the individual engines in the BOS approach, which will include a comparison of the results of our system with human judgments and subsequent modification of the system based on this feedback. We also intend to experiment with a training schedule by which the dispatcher could be trained over text samples, by trying potentially random assignments of text parts to modules and then seeing which assignment regimes produce the best results. A variant on this would be human text "tagging" by intuitions about the text type (where the human tagged it by the module type that he considered would be needed; this would be essentially a difficulty rating the text a priori), and again assessing this against system results. As the size of such an experiment can be quite significant, we envisage the use of some form of quasi-automatic quality scoring for MT of the sort proposed recently by Henry Thompson and his colleagues (e.g., Brew and Thompson, 1994).

References

Ben Ari, D. M. Rimon and D. Berry. 1988. Translational Ambiguity Rephrased. Proceedings of the Second International Conference on Theoretical and methodological Issues in Machine Translation. Pittsburgh, June 1988.

Brew, C. and H. Thompson. 1994. Automatic Evaluation of Computer Generated Text: A Progress Report on the TextEval Project. Proceedings of HLT94, Princeton, March.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R.L. and Roossin P.S. 1990. A statistical approach to language translation, *Computational Linguistics*. vol 16, 79–85.

Furuse, O and H. Iida. 1992. An Example-Based Method for Transfer-Driven Machine Translation. Proceedings of TMI-92. Montreal. 139-50.

Grishman, R. and M. Kosaka. 1992. Combining rationalist and empiricist approaches to machine translation. Proceedings of TMI-92. Montreal. 263-74

Isabelle, P. and L. Bourbeau. 1985. TAUM-AVIATION: Its Technical Features and Some Experimental Results. *Computational Linguistics*, 11: 18-27.

Jones, D. 1992. Non-hybrid example-based machine translation architectures. Proceedings of TMI-92. Montreal. 163-71.

Lehmann, H. and N. Ott. 1992. Translation relations and the combination of analytical and statistical methods in machine translation. Proceedings of TMI-92. Montreal. 237-48.

Maruyama, H. and H. Watanabe. 1992. Tree cover search algorithm for example-based translation. Proceedings of TMI-92. Montreal. 173-84.

McLean, I. 1992. Example-based machine translation using connectionist matching. Proceedings of TMI-92. Montreal. 35-43.

Nirenburg, S., J. Carbonell, M. Tomita and K. Goodman. 1992. **Machine Translation: A Knowledge-Based Approach**. San Mateo, CA: Morgan Kaufmann.

Su, K-Y. and J-S. Chang. 1992. Why corpus-based statistics-oriented machine translation. Proceedings of TMI-92. Montreal. 249-62.

Wilks, Y., Fass, D., Guo, C-M., McDonald, J., Plate, T., & Slator, B. 1990. Providing Machine Tractable Dictionary Tools. *Machine Translation*, 5:2, 99-151.