THE BKB (BILINGUAL KNOWLEDGE BANK): A PROMISING NEW APPROACH TO MACHINE TRANSLATION

Claude Bédard
TRADUCTIX

Keywords: Machine Translation, Interlingua, Parsing, Natural Semantics, Dependency Grammar, Bilingual Knowledge Bank, Translation Unit, Implicit Knowledge

Abstract: The DLT machine translation system has many original features: it uses an interlingua (Esperanto) and dependency (rather than constituency) grammar. It is designed for use by monolinguals by means of a disambiguation dialogue. A more recent feature is a new data structure called BKB, which represents lexical, syntactic, and semantic information implicitly. The BKB plays a prominent role in parsing, lexical choice and target language idiomaticity. This approach, based on the "translator's experience", seems quite promising.

1. MAIN FEATURES OF THE DLT SYSTEM

The DLT (Distributed Language Translation) system has been developed by the Dutch software firm BSO since 1984, and is planned to come on the market in the course of the 1990's. I have recently been invited to Utrecht to study the system. Here is a "nutshell" description of its main features and technological options.

DLT is an interlingual system, i.e., a system which goes through an interlingua for translating from and into many source and target languages. (A well-known advantage of interlingual systems is the reduction in the number of transfer modules between the languages involved.) The chosen interlingua is Esperanto, a planned human language which has several interesting features, the main ones being that it is reasonably free of homonymy and syntactic ambiguity, and that it has a very regular, agglutinative morphology which enables using morphemes as so-called "semantic primitives".

Corollary to the choice of a human language as an interlingua, the resulting translation from source language to target language involves in fact two successive translations, the first into Esperanto, and the second into the selected target language.

DLT uses dependency grammar, as opposed to the more conventional constituency grammar, for syntactic representation. The two advantages of dependency grammar are that it is more universal (constituency grammar is rather suited to languages such as English, which has a rather fixed word order) and also better describes role-based relations between the words; as such it is better suited to semantic processing.

DLT tackles head-on the problem of lexical choice (the choice between multiple equivalents of a given word), often sidestepped by conventional MT systems. The tool used for this purpose had been the LKB (Lexical Knowledge Bank), a collection of typical, short semantic contexts, which has now evolved into the more powerful concept of BKB (Bilingual Knowledge Bank). This has become the salient feature of the DLT system, which I will describe in more detail in the following sections – although it has not been fully implemented yet.

Last but not least, DLT is designed to be used primarily by monolinguals knowing only the source language. The user is requested to answer questions asked by the system to disambiguate the source language. These questions are asked in the source language only.

2. THE BILINGUAL KNOWLEDGE BANK

The BKB concept stems from the need to encode a vast amount of complex linguistic information into any real-life MT system aiming at good quality output.

2.1 INFORMATION NEEDED

First, lexical information is needed, giving the range of alternative translations for a word, as well as criteria for choosing between these. In addition, phrases are required in the cases where words cannot be translated in isolation, as well as any syntactic transformation rules required by the equivalent in the target language.

Syntactic information is necessary both to analyze or parse the SL (source language) sentence and to effect whatever structural transformations are necessary in the TL (target language). So far, it has proved very difficult to formalize adequately the syntax of a language, with all its endless subrules and exceptions.

Information about idiomaticity is also required. It is frequent that a rather literal translation, even with no grammatical errors, is felt as quite "unnatural". How do we model this kind of information: detecting unnnaturalness and triggering the appropriate changes? This problem has not been solved yet.

Semantic information is necessary to guide syntactic and lexical choices. Coding this information typically involves first designing a set of semantic codes, and then applying them coherently on all lexical entries and selection rules. This has proved to be full of pitfalls and to have still a limited effectiveness.

One problem which comes out strongly from these desiderata is that expressing all this knowledge explicitly is time-consuming and, what is more, requires a very high degree of skill; so that coding a complete set of information for a rather high-quality translation is long and very costly, and somewhat fragile and error-prone.

2.2 KEEPING IT IMPLICIT

The BKB proposes an interesting solution to these shortcomings by using implicit (rather than explicit) knowledge. Its primary material is a "bitext", or corpus of texts in two languages. On one side of the BKB is the English or French or German-language version of the corpus, and on the other its Esperanto translation. Either side of the BKB can be used as source or target language.

In order to be machine-useable, this bitext requires a high degree of structuring. First of all, the sentences of each monolingual text are parsed to form dependency trees. Next, these trees are linked together from one language to the other; this linking is not word-for-word, but is based on translation units. Lastly, conceptual links are established between various content words within each text.

Translation Units

In the BKB, a translation unit (TU) is defined as any two portions of sentence which are considered equivalent. The concept of TU is necessarily bilingual, and each TU is defined in terms of the two languages involved.

TUs may relate either individual words or groups of words; their chief interest is that they may relate groups of words which are not symmetrical translations of each other, although translationally equivalent. Let us consider the following sentence (for easier

understanding, the languages used in the examples do not include Esperanto) :

> The board of PAC unanimously confirms the mandate.
> Le conseil du PAC est unanime dans sa confirmation du mandat.

Here, unanimously confirms is related to est unanime dans sa confirmation de; this TU cannot be further subdivided and is therefore considered minimal. Here we see that the BKB offers much more than a conventional MT dictionary, which is more or less restricted to a limited range of lexicalized translations. The BKB offers "organic" solutions drawn from real texts.

TUs can be used in generalized ways, and include structural information. For example, if the following sentence is present in the BKB:

> In the pages which follow some relevant examples are given.
> Les pages qui suivent présentent quelques exemples pertinents.

The structural transformation from one language to the other is captured by a TU linking the two dependency grammar trees; this TU can be applied to produce the same effect in various other situations, such as:

> In the next chapter a case study is given.
> More explanations are given in this section.

These few examples show that TUs can formalize a rather elusive but well-known aspect of translation expertise: skewing, where the translator builds an asymmetrical translation. The BKB is a good starting point for automatically deciding when skewing is necessary, and how to word this skewing.

Natural semantics

In the DLT system, semantics is kept implicit; it is not based on a set of codes, but on contexts, of which the BKB provides a rich supply. Let us consider the following passage.

> The board of PAC unanimously confirms the mandate. Moreover, since its members so easily reached a consensus, the board considers that such decisions and other related rulings should be left to the local committees (e.g. LSC or RREC). It also approves...

The first elements of semantic knowledge are the basic syntactic associations between words, obtained by the parsing process. For example, we can deduce from the first sentence that board is a possible subject for confirm (and conversely), mandate a possible object for confirm, etc., when we need this type of information in future situations.

Then, some basic semantic features (IS-A, PART-OF, HAS, etc.) can be implied from the text. In the example:

- The words decision and rulings are coordinated, therefore are considered somehow compatible semantically.
- The word e.g. implies that LSC and RREC are instances or members of the set committee.
- The pronoun its allows to infer that a board has such things as members.

Finally, coreferential links are drawn to equate board of PAC, board and it; this allows pooling the various clues gathered by each instance, and enriches the semantic picture of the concept.

Building the BKBs

Building a BKB is no small task, of course. First the corpus must be available in several languages, including Esperanto. The building process is an interactive one, called synsemization, where a human operator answers the machine's questions or validates its choices. Compared with conventional MT coding, synsemization is expected to be less labour-intensive and also to require less theoretical expertise since the work consists in applying factual judgements on existing data rather than adding new data from a formalized system of artificial rules and features (with all the hazards of subjectivity and undecidability inherent in such an approach). The machine is also expected to take on an increasingly large portion of the work as its knowledge increases.

3. TRANSLATING WITH BKBs

Within the interlingual framework of DLT, two successive translations are involved: one from the SL to Esperanto, and another from Esperanto to the TL. Each of these steps uses a separate BKB. Here is a very brief overview of the process involved in each translation.

The parser contains the morphological reduction rules and the basic constraints on combining syntactic units in the SL. Further processing is based on consulting the BKB, using pattern-matching techniques; the parser tries to match tentative structures onto the trees already present in the BKB. In fact, the BKB implicitly describes the grammar of the language by offering a collection of reference parse trees. The parsing process is therefore probabilistic, and can achieve more relevant results faster by preferring what has been already experienced over what has not.

The choice between the alternate TL equivalents is done, like the parsing, by pattern matching. Preference is normally given to longer matches between the contexts of the sentence being translated and the contents of the BKB; frequency and recency are other factors. This pattern matching enables to retrieve not only individual words, but also groups of words of any length.

During these two processes, a disambiguation dialogue initiated by the machine helps make the right choices.

Finally, the TL tree fragments are assembled into a sentence, with various checks for coherence and well-formedness in the TL grammar. Once a sentence has been translated, it is incorporated into the BKB and is available for processing the subsequent sentences.

4. THE VALUE OF THE BKB CONCEPT IN MACHINE TRANSLATION

The BKB is a hypertextual, multi-level data structure where lexical, syntactic, and conceptual information, whether contrastive or language-specific, are all accessible at any time. This fascinating reservoir of knowledge is a radical departure from conventional data structures for MT.

The BKB concept is based on the intuition that since the task to be done is one of translation after all, the most relevant source of information could well be the "trace" left by the translator: the translated version, along with the source text. By using actual translation as its main knowledge resource, the DLT system comes closer than any MT system I know to applying a translation model rather than doing linguistic transcription.

The DLT system has been based on the idea that human language is an acceptable — and robust — means of representing knowledge; hence the choice of Esperanto as an interlingua (over purely artificial interlingual representations, based on cognitive techniques and semantic coding, which have so far proved too fragile outside strictly controlled environments). Again this principle is reaffirmed in the BKB. After decades of possibly overformalized explicit linguistic description, this looks like an exciting new avenue to explore.

The information present in the BKB does not allow for deep explicit understanding of the text. But we should realize that, contrary to text-understanding applications, machine translation can be achieved in most cases with a moderate degree of understanding; what is important in translation is to select between alternatives, and this can largely be done by an accumulation of low-level clues and inferences. Also, let us not forget that MT needs much more than cognitive information; information about linguistic usage and translation expertise are directly relevant to the task.

Finally, one great advantage of implicit human-language knowledge as data is that such a representation is not committed to a particular cognitive theory. The BKB is also independent from the processing mechanism itself; future improvements or changes to the processing rules will not alter the validity of its contents.

I do not like to use the word "breakthrough", which has been so much overused in a field in which breakthroughs have been rare indeed. But the BKB concept seems to me a major innovation and a very promising avenue for MT. There are still many challenges on the way to harnessing the "natural resource" of implicit knowledge, but the potential of the approach is obvious.

5. REFERENCES

1. Papegaaij, Bart and Schubert, Klaus. Text Coherence in Translation. Foris Publications, Dordrecht (Holland)/Providence (RI, USA), 1988.

2. Sadler, Victor. The Bilingual Knowledge Bank - A New Conceptual Basis for MT. BSO, Internal Publication, March 1989.

3. Sadler, Victor. Translating with a Simulated Bilingual Knowledge Bank (BKB). BSO, Internal Publication, April 1989.

4. Sadler, Victor. 28 Questions and Answers about the BKB (Bilingual Knowledge Bank), the Framework for Industrial Production of DLT. BSO, Internal Publication, April 1989.

5. Sadler, Victor. Analogical Semantics. Foris Publications, Dordrecht (Holland)/Providence (RI, USA), 1989.