STS: An Experimental Sentence Translation System

Eric Wehrli*
University of Geneva
1211 Geneva 4
wehrli@uni2a.unige.ch

Abstract

STS is a small experimental sentence translation system developed to demonstrate the efficiency of our lexicalist model of translation. Based on a GB-inspired parser, lexical transfer and lexical projection, STS provides real-time accurate English translations for a small but non-trivial subset of French sentences.

1 Introduction

Accurate natural language translation requires a wide range of cognitive abilities, including grammatical knowledge of the languages involved but also some amount of extralinguistic knowledge and common sense reasoning. Lack of well-defined theoretical models of knowledge and common sense reasoning makes fully automatic high quality translation unlikely in the near future.

In the meantime, what appears to be an increasingly appealing alternative is on-line interactive machine translation, *i.e.* systems which can consult the user when they are unable to solve a problem[1]. However, in order to be a viable alternative to other machine or machine-aided translation models, and in addition to the usual requirements of reasonable quality and low cost, an on-line interactive system must also satisfy the requirements of real-time systems and in particular be fast enough not to use the user's patience.

*Part of the work described in this paper has been supported by a grant from the Swiss national science foundation (grant no 11-25362.88).

[1]For a discussion of on-line interactive translation see Kay (1982), Tomita (1986), Johnson and Whitelock (1987), among others.

As a preliminary step towards the development of an on-line interactive system, we develop the STS system to establish whether our lexicalist translation model, using a GB-inspired parser, lexical transfer and lexical projection, provides the kind of efficiency required for on-line translation[2]. In its present implementation, STS translates sentences from French to English. It can handle a limited, although not trivial subset of clausal structures, using a lexical database of more than 5,000 entries (including compounds and idiomatic expressions).

2 Architecture of the system

The basic architecture of the STS system is the familiar transfer model with its three main components: analysis, transfer and generation. To see how these components interact, consider the following (drastically oversimplified) sketch of the translation process: an input sentence in the source language (SL) is first mapped onto some formal representation for this language (S-structure). This is done by the parser, on the basis of lexical information and detailed knowledge about the grammar of SL. The transfer component maps then the S-structure returned by the parser onto an appropriate S-structure in the target language. The transfer is done constituent by constituent, in a top-down fashion, starting with the top S (or $\overline{S}$) constituent. For each constituent, the lexical head is first considered: Its lexeme is associated with a set of possible translations, *ie.* one or more lexemes in the target language. Once the relevant lexeme has been selected an appropriate structure is projected on the basis of its lexical properties

[2]For a description of the parser used in the STS project, see Wehrli (1988).

and of the general rules and principles of target language grammar. Notice that the projection is done solely on the basis of information internal to the target language.

In other words, the interface structure is minimal, as it should be, and is almost entirely a matter of lexical mapping. Both the analysis and generation modules are completely independent of each other. This again, is a desirable feature, in the sense that, for instance, the same parser can be used no matter what the target language might be. In fact, given that the S-structures in this system are solely justified in terms of the grammar, the parser (and generator) do not have to be application dependent.

2.1 Lexical database

The lexical database is the central piece of the STS system. It contains crucial information used by the three active components of the system. This information is distributed in two monolingual lexicons (SL lexicon and TL lexicon) along with one bilingual lexicon. We shall consider them in turn:

2.1.1 Monolingual lexicons

We assume a static – or relational – conception of morphology, along the lines of Jackendoff (1975), Wehrli (1985). According to this view, morphological relations between two or more lexical entries are expressed by a complex network of relations.

A monolingual lexicon distinguishes three basic entities: *lexeme*, *word* and *idiom*. A lexeme is an abstract lexical unit, which can be compared, roughly speaking, to a standard dictionary entry. It stands for a whole class of morphological variants. By contrast, a *word* corresponds to a particular morphological instantiation of a lexeme. In other words, we make a clear distinction among features which may vary with inflexion and those which are invariant. To give an example, *am*, *are*, *were*, *being*, *be*, etc. are words, morphological variants of the lexeme "be". The lexemes are associated all the features which are independent of the morphological realization, such as semantic features, subcategorization features, and the like. Features which depend on inflexional markers – *e.g.* tense, number, person, *etc.* – are naturaly attached to the words. In addition to words and lexemes, a monolingual lexicon also contains a list of idioms, *ie.* pharses which have a fixed, non-compositional meaning, such as *to kick the bucket* or *to be caught red-handed.*

The notion of lexeme turns out to be one of great significance: Not only does it make possible to factor out basic syntactic and/or semantic properties shared by morphologically related words. At the same time, it also provides the abstract lexical level which is relevant for lexical transfer.

2.1.2 Bilingual dictionary

The bilingual dictionary specifies the set of possible relations between lexemes of the source language and lexemes of the target language. Each entry in this dictionary specifies one SL lexeme and one TL lexeme. In case one particular SL lexeme has more than one corresponding TL lexeme (e.g. *aimer* $\longrightarrow$ *to like*, *to love*, etc.), the bilingual dictionary contains as many entries as there are correspondences. The bilingual dictionary contains other kind of information as well. For instance, in the case of argument-taking elements, such as verbs or predicative adjectives, an entry of the bilingual dictionary must also specify how the arguments of the SL predicate match the arguments of the TL predicate.

3 The transfer component

The role of the transfer component is to map SL S-structures onto TL S-structures. In STS, this mapping is done indirectly, by means of two mechanisms: lexical transfer and lexical projection.

Transfer applies to the syntactic structures returned by the parser. In a top-down fashion, starting with the main S-structure, the transfer procedure considers the lexical head of a phrase, look it up in the bilingual dictionary and selects the most appropriate TL lexeme, based on contextual information, features in the bilingual dictionary. Once a lexeme has been selected, a process of lexical projection creates a TL syntactic structure on the basis of the lexical properties of the TL lexeme, and of the general syntactic

properties of TL. In the next step, the transfer procedure considers the complements of the head, using the same strategy. In addition to lexical transfer and lexical projection, the transfer procedure is guided by the argument mapping information found in the bilingual dictionary, as mentioned above.

In the STS system, discrepancies between SL and TL, such as differences in word order or argument matching, can be handled quite naturally without the need of complex and *ad hoc* structural transfer rules. To illustrate this point, the fact that a sentence containing a modal verb must be assigned a bi-sentential structure in French, but not in English follows from the lexical properties of modal verbs in French and in English, *i.e.* French modals are main verbs selecting an infinitival sentential complement, while English modals are not marked as main verbs. Within a linguistic theory which assumes that phrase structures are projected from their lexical elements, the structural differences between French and English sentences follows from the lexical differences between, say, *pouvoir* and *can.*

4 Some examples

Two examples of translations produced by the STS system are given below. In the first example, *a* is the input sentence, *b* the structure returned by the parser, and *c* the sentence returned by the translator[3]. The structure has been omitted in the second example.

(1)a. le livre que vous regardez semble être facile à lire.

b. [$_{S}$ [$_{NP}$ [$_{NP}$ le livre]$_j$ [$_{\overline{S}}$ [$_{COMP}$ [$_{NP}$ que]$_j$] [$_{S}$ [$_{NP}$ e]$_k$ [$_{VP}$ [$_{CLI}$ vous$_k$] regardez [$_{NP}$ e]$_j$]]]]$_i$ [$_{VP}$ semble [$_{\overline{S}}$ [$_{S}$ [$_{NP}$ e]$_i$ [$_{VP}$ être [$_{AP}$ [$_{NP}$ e]$_i$ facile [$_{\overline{S}}$ à [$_{S}$ [$_{NP}$ e] [$_{VP}$ lire [$_{NP}$ e]$_i$]]]]]]]]]

c. the book at which you are looking seems to be easy to read.

(2)a. l'homme dont vous semblez avoir oublié le nom ne pourra-t-il pas vous fournir les renseignements dont vous avez besoin?

b. won't the man whose name you seem to have forgotten be able to provide you with the information that you need?

Such examples show that the STS system can succesfully handle structures of a non-trivial level of complexity. The second example, in particular, shows the ability of this system to handle problems such as difference in word-order, argument matching and idiomatic expressions. In addition, this model proved extremely efficient – total time for parsing and translation of the above sentences averages 150 ms/word[4] – which is a crucial prerequisite for on-line interaction.

References

Jackendoff, R. (1975). "Morphological and semantic regularities in the lexicon," *Language* 51:639-671.

Johnson, R. and P. Whitelock (1987). "Machine Translation as Expert Task," in S. Nirenburg (ed.) *Machine Translation*, Cambridge University Press.

Kay, M. (1982). "Machine Translation," *American Journal of Computational Linguistics* 8:74-78.

Tomita, M. (1986). *Efficient Parsing for Natural Language*, Kluwer Academic Publishers.

Wehrli, E. (1985). "Design and Implementation of a Lexical Data Base," *Proceedings of the 2nd European ACL Conference*, pp. 146-153.

Wehrli, E., 1988. "Parsing with a GB grammar," in U. Reyle and C. Rohrer (eds.), *Natural Language Parsing and Linguistic Theories*, Reidel, 1988.

[3]The structures returned by the parser correspond to slightly simplified S-structure representations of a GB-grammar. The indexes in (1b) express $\overline{A}$-binding or control relations with empty categories (*e.g.* [$_{NP}$ e]).

[4]CPU time on a DEC Vaxstation II computer.