Translation of Words in Context

Eric Wehrli
LATL-Department of Linguistics
University of Geneva
Eric.Wehrli@lettres.unige.ch

Abstract

TWiC is an on-line word and expression translation system which uses a powerful parser to (i) properly identify the relevant lexical units, (ii) retrieve the base form of the selected word and (iii) recognize the presence of a multiword expression (compound, idiom, collocation) the selected word may be part of. The conjunction of state-of-the-art natural language parsing, multiword expression identification and large bilingual databases provides a powerful and effective tool for people who want to read on-line material in a foreign language which they are not completely fluent in. A full prototype version of TWiC has been completed for the English-French pair of languages.

1 Introduction

In the information technology age, information is expected to be readily available, for instance on the Internet. However, in a multilingual world, it is often the case that the available information happens to be in a language other than the user's native language. For instance, a huge number of students and scientists around the world routinely browse web pages in English. While many of them have a sufficient knowledge of English to retrieve the information they need, a significant number of them may experience terminological difficulties.

Available on-line machine translation tools, such as *Babel Fish* on the *AltaVista* search engine site, may prove useful for those users with little or no knowledge of the relevant language. However, given the current state of machine translation, the risk of inaccurate translation, or even of clumsy translation, is such that most users who are fairly fluent in the relevant language, but with some gaps in the terminology, would much rather read the document in its original form, if they had an adequate terminology help at hand.

Current on-line bilingual terminology databases can satisfy this need (For instance Eurodicautom, the European terminology database or, for French-English, *Le grand dictionnaire terminologique de l'Office québécois de la langue française*, among others.), but only in part, as they suffer from the following shortcomings:

- On-line bilingual database lookups tend to be ``noisy", in the sense that they are likely to return a large number of translations for a given word, many of them irrelevant in the given context.
- Easy access (ie. clicking or selecting a word or an expression) will only be possible for words which occur in their base form.
- Multiword expressions (MWEs) are poorly handled. Although many expressions and collocations are listed in good bilingual dictionaries, they may be difficult to find, especially when the user only selects one particular chunk. For instance, consider the collocation *school of fish*. If you select school, a good dictionary is likely to let you read through several pages before mentioning this particular sense of the word.

The second shortcoming (lack of lemmatization) and very partially the first one are addressed by more sophisticated natural language processing systems, such as the ones commercialized by Babylon and Xerox corporations, but to the best of our knowledge, neither of them can handle non adjacent collocations in a satisfactory fashion.

2 The TWiC System

TWiC (Translation of Words in Context) is an on-line word and expression translator which uses a linguistic parser and other computational tools to better select the proper lexical unit(s) to translate, i.e. to drastically reduce the "noise" of word translation. When clicking on a word, a detailed linguistic analysis of the whole sentence is performed, which (i) provides the base form of the selected word, (ii) eliminates all the readings which are not syntactically compatible with the linguistic environment, and (iii) identifies MWEs such as compounds, idiomatic expressions and collocations, the word might be part of.

For instance, when the word *rose* in sentence (1) below is selected, TWiC identifies the verb *to rise* as the proper lexical item and, accordingly, displays *se lever* ou *s'élever* rather than the translation linked to the adjectival or nominal readings of *rose*.

(1) They all rose.

2.1 Multiword expressions

Perhaps the most original feature of TWiC is its ability to recognize that a selected word is part of a multiword expression and to give the appropriate translation for that expression. Consider, for instance, the following examples, where the word in boldface corresponds to the selected item.

(2)a. Just then a **school** of little fishes swam past.
b. A new record ... has been **broken**.
c. He **gave** up his leadership.

In sentence (2a), TWiC identifies the collocation *school of fish* and provides the proper translation *banc de poissons*. Sentence (2b) illustrates the case of a "verb-object" collocation, *break-record*, where the chunks can be several words apart (and not necessarily in the expected order). The parser can identify this collocation and TwiC suggests the translation *battre-record*. The word *gave* in sentence (2c) is part of the lexical item *give up*, and TWiC displays *abandonner* ou *renoncer à* rather than the several dozens of possible translations of the word *give*.

2.2 Architecture of TWiC

The TWiC system is composed of the following components:

- A sentence extractor, which attempts to extract the sentence containing the word selected by the user, on the basis of document formatting as well as typographical clues.
- A linguistic parser, which performs a full linguistic analysis of the sentence, in order to disambiguate and more generally narrow down the relevant lexical item selected by the user.
- A large bilingual database containing correspondences for words, compounds, collocations and idiomatic expressions.
- A graphical user interface, capable of displaying the result(s) of a request in the form of hypertext link. If the link is active, the user can click on it to see other suggestions.

The linguistic parser is the key component of TWiC in the sense that it is by the application of the parser's morphological and syntactic filters that the relevant lexical item can be selected. This parser, which is based on an adaptation of Chomsky's (1995) linguistic theory, has been described in details in Wehrli (1997). For the TWiC application, it returns the following information for the selected item:

- A morpho-syntactic label specifying its word class along with inflectional features.
- Its position in the sentence.
- Its base form (more precisely the lexeme number corresponding to its base form).

On the basis of that information, TWiC retrieves the available translation for the appropriate lexeme from the bilingual database. If the lexeme corresponds to a multiword expression, the whole expression is retrieved and displayed to the user.

Notice that in order to correctly handle multiword expressions, superficial syntactic tagging would not suffice. What is needed is a more comprehensive analysis, capable of interpreting extraposed elements such as so-called *wh*-interrogative phrases, relativized phrases, subject of passive verbs, etc. Consider, for instance, the example (3) below of a noun phrase modified by a relative clause.

(3) The record she was hoping to be able to break (...).

In this example, the head noun *record* is modified by the relative clause *she was hoping to be able to break*. Since the unexpressed relative pronoun corresponds to the direct object of *break*, we have an occurrence of the "verb-object" collocation *break-record*. In order to reach this conclusion, the parser must be able (i) to recognize the presence of a relative clause, (ii) to identify the antecedent of the relative pronoun, and (iii) to establish a link between this antecedent and the verb on which the relative pronoun syntactically depends. As illustrated by the structure (4b) produced by the parser for the slightly simplified version of our last example given in (4a), the TWiC parser does exactly that. A chain is established, first between the canonical direct object position (represented by the empty element $[_{DP}$ e $]$, which stands for the trace of the extraposed element) of the verb *break* and the (abstract) relative pronoun, and then between this pronoun and its antecedent *record*. In the structures returned by the parser, chains are expressed by means of coindexation. The double chain connecting first the noun *record* and the (abstract) relative pronoun in the specifier position of CP, and then the trace of the relative pronoun in the direct object position of the verb *break*, is represented in the structure (4b) by the index ***i***.

(4)a. The record she might break...
b. $[_{DP}$ the $[_{NP\,i}$ record $[_{CP}$ $[_{DP}$ e$_{i}$ $]$ $[_{TP}$ $[_{DP}$ she $]$ might $[_{VP}$ break $[_{DP}$ e$_{i}$ $]]]]]]$

An illustration of a partial output of the parser for the fragment *They foiled an attempt(...)* is displayed in the following table:

Source word	POS tag	Position	Lexeme number
they	PRO-PER-3-PLU	0	111000011
foiled	VER-PAS-3-PLU	5	141000136
an	DET-SIN	12	111050002
attempt	NOU-SIN	15	111005034

For each word, the parser returns a POS tag (morpho-syntactic label), such as PROnoun, VERb, DETerminer and NOUn, along with inflectional features, such as PERsonal, PASt, SINgular, PLUral. The third column specifies the position of the word in the sentence. The lexeme number, in the last column, refers to the unique internal number of a lexical item in our lexical database.

The need for syntactic normalization in order to properly identify particular types of collocations is not restricted to the case of relative clauses, as argued in Wehrli (2000) and Goldman *et al.* (2001). Similar arguments can be made for all cases of extraposed elements, including passives, topicalized phrases, and raising structures.

3 Current status

An advanced prototype of the TWiC has been completed for the English-French language pair with a bilingual database of over 45'000 correspondences, including 1'500 collocations. The current version runs under MS-Windows. It is planned to develop an Internet version of TWiC (in the form of a plug-in) and to extend it to other languages, including German and Spanish.

4 Acknowlegments

Thanks to Luka Nerima Anne Vandeventer Faltin, Catherine Walther Green and Christopher Laenzlinger for comments on an earlier version of this paper. This research was supported in part by a grant from the RUIG/GIAN organization.

5 Bibliographical References

Noam Chomsky. 1995. *The Minimalist Program*, Cambridge, MIT Press.

Jean-Philippe Goldman, Luka Nerima and Eric Wehrli. 2001. Collocation Extraction Using a Syntactic Parser. In *Proceedings of the ACL*, pp. 61-66.

Eric Wehrli. 1997. *L'analyse syntaxique des langues naturelles: problèmes et méthodes*, Paris, Masson.

Eric Wehrli. 2000. Parsing and Collocations. In D. Christodoulakis (ed.) *Natural Language Processing-NLP 2000*, Springer Verlag, pp. 272-282.