Jacques Ladouceur, Arman Tajarobi, Université Laval, Canada

Computer-Assisted Plurilingual Reading System

Due to the common roots of many European languages such as French, English, German, Spanish, Portuguese, Italian, etc., many formal similarities can be found in these languages in the nominal category. In addition, a great number of technical terms are neologisms created in one language and borrowed by other languages with very little modification at the morphological level. These two facts may be exploited to permit the extraction of equivalent terms from parallel texts.

Pluriterm is a tool that uses the results of Termplus, which is a term extraction unit capable of retrieving an extremely high percentage of complex terms (nearing 99%) in a text. With the recent addition of the Neutraliser, Termplus is now able to retrieve not only almost all of the complex terms in a text, but also a very high percentage of simple terms.

The base technology of parallel text alignment as exploited by Pluriterm is very simple and straightforward. It consists in using the formal properties of a term in language A to produce a list of graphically similar terms in language B. If the graphical similarity of terms can not be exploited directly, two other techniques are used. The first is the use of conceptual networks. This technique is based on the premise that complex terms containing one or more graphically similar units are conceptually related. For example, there is a conceptual relationship between local drive, built-in drive, disk drive and CD-ROM drive, where the element drive is common to all terms. This conceptual relationship can be exploited as an anchor to go from a term TA in language A to a term TB in language B even if there is no formal resemblance between terms TA and TB. In a recent experiment using the conceptual network of lecteur interne in French, we arrived at built-in drive in English using internal drive as the bridge between the two terms:

- lecteur interne
- internal drive
- built-in drive

The second technique is the use of collocations to establish bridges between two terms, TA and TB, by exploiting the graphical similarities of the contextual environments of TA and TB. The technique is based on the premise that there is a conceptual bond between terms occurring near each other in a text. For example, speaker and sound card are more likely to be used in the same sentence or paragraph than are speaker and local hard drive. Since there is no formal resemblance between speaker and the French equivalent haut parleur, the collocations of these terms may be used to establish a link between these terms in parallel texts. Sound card and carte de son, which are graphically close, are likely to appear at about the same frequency and average distance from speaker and haut parleur in two large sets of parallel texts. Similarly, microphone, which is graphically similar in both French and English, has a better chance to appear more frequently and at closer distances to speaker or haut parleur than to power supply or mother board.

- speaker
- microphone
- haut parleur
- sound card
- carte de son

The case of Pluriterm demonstrates that very simple formal manipulations may be used to handle efficiently domain-related concepts without having to deal directly with the complexities of meaning and semantics. The ideas forwarded here may also be exploited for other text manipulation tasks. In all cases, the objective is to enable the machine to read in a non-linear manner by targeting specific information in large masses of documents.