TALN 2010, Montréal, 19–23 juillet 2010

Moz: Translation of Structured Terminology-Rich Text

Graham Russell
Onscope Group Inc., 651 Notre-Dame Ouest, Montréal QC, Canada H3C 1J1
grussell@onscope.com

Résumé. Description de Moz, un système d'aide à la traduction conçu pour le traitement de textes structurés ou semi-structurés avec une forte proportion de contenu terminologique. Le système comporte une mémoire de traduction collaborative, qui atteint un niveau élevé de rappel grâce à l'analyse sous-phrastique ; il fournit également des dispositifs de communication et de révision. Le système est en production et traduit 140 000 mots par semaine.

Abstract. Description of Moz, a translation support system designed for texts exhibiting a high proportion of structured and semi-structured terminological content. The system comprises a web-based collaborative translation memory, with high recall via subsentential linguistic analysis and facilities for messaging and quality assurance. It is in production use, translating some 140,000 words per week.

Mots-clés : Aides à la traduction, sous-langage, analyse conceptuelle.

Keywords: Translation aids, sublanguage, conceptual analysis.

1 Introduction

This article describes Moz, a translation support system designed for high-volume processing of structured terminology-rich text. By this we mean texts such as catalogues, customs declarations, descriptions of products and services, and so on. Documents of this kind typically contain a large proportion of term-like nominal phrases arranged in list structures and linked by expressions representing a relatively restricted range of semantic relations. An example is shown below :

> Métaux précieux et leurs alliages autres qu'à usage dentaire ; joaillerie, bijouterie, pierres précieuses, horlogerie et autres instruments chronométriques nommément montres et chronomètres ; cuir et imitations du cuir ; peaux d'animaux ; malles et valises ; parapluies, parasols et cannes ; fouets et sellerie ; vêtements nommément chemises, pantalons, lingerie de corps, manteaux, gants, mitaines, costumes, costumes de mascarades, vestes, bas, chandails ; chaussures (à l'exception des chaussures orthopédiques) nommément bottes, sandales, souliers, chaussures de sport, souliers de gymnastique, mocassins, sabots ; chapellerie nommément chapeaux, bérets, casquettes visières, bonnets.

Such texts are sufficiently homogeneous in style and syntax to permit sublanguage approaches to linguistic analysis.

2 The System

2.1 Background

Moz is currently in production use under a contract from the Canadian Intellectual Property Office (CIPO) for the translation of trade-mark descriptions between English and French. The relevant text describes the areas of activity (products and services) for which the mark is claimed, and has the form illustrated above.

The system has been in operation since February 2007, with almost 20 million words translated to date. Currently, approximately 140,000 words are translated per week ; some 1,000 new marks are filed each week, and 600 modified. The fact that a third of the input takes the form of modifications to existing marks makes the translation memory approach (section 2.4) particularly advantageous.

The system is bidirectional, translation being performed both from English to French and from French to English. The former direction dominates in terms of volume, however. Moz is web-based and accessible to remote users.

2.2 Work Flow

The translation process proper follows a conventional pattern :

1. Text analysis : input text is segmented and analysed.
2. Pretranslation : where available, each source-language segment is paired with the best candidate translation drawn from the translation memory.
3. Translation : the translator considers alternative translations from the memory or provides a new one where appropriate.
4. First revision (terminology)
5. Second revision (structure)
6. Modification of the memory : any new equivalents or changes in status are recorded.

Only the first of these steps, described in section 2.3, differs significantly from the normal context of translation memory use.

Additional facilities are provided for project management (section 2.5.1) and communication between members of the translation team and revisers (section 2.5.2).

2.3 Text Analysis

Input text is analysed in greater detail than is normal with commercially deployed translation memories. Couturier *et al.* (2006) give an overview of some of the methods employed.

Briefly, a mixture of symbolic and statistical techniques are applied in a fine-grained segmentation process which identifies candidates for minimal translation units. The latter are typically single words or short multiword expressions, and often correspond to independently motivated terms. Sample results of this stage of the process can be seen in the source-language column of figure 1. Additionally, certain semantic

HYPONYM	Produits alimentaires **nommément** marinades, oeufs, bouillons, ...
USER	Medicated tooth whitening kit **for use by** dental professionals
DOMAIN	Computer software **for use in the fields** of transportation management and supply chain management
NEGATION	Machines à écrire et articles de bureau (**à l'exception des** meubles)

TAB. 1 – Some semantic relations and associated lexical cues

relations between segments are detected, based on the presence of lexical-syntactic patterns ; table 1 gives some examples.

The same approach underlies other applications in the trademark domain at Onscope : text classification and filtering of search results both exploit the detailed information thus obtained.

2.4 Translation Memory

The translation memory contains entries for terms and other expressions occurring in trademark texts. In use, the translator is presented with the result of an automatic pretranslation phase, in which source-text segments have where possible been assigned candidate translations. Figure 1 shows a typical configuration.

In practice, pretranslation recall is often very high, for a number of reasons. The input texts are by their nature repetitive and formulaic ; moreover, as noted in section 2.1, a significant portion of the input consists of revisions to previously translated texts, whose content has therefore already been stored in the memory. In addition, the fact that the segments involved are typically much shorter than the clauses forming the basis of most translation memories allows for less variation and thus makes retrieval of an exact match more likely ; even when an entire sentence is not present in the memory, some of its components may be.

Alternate translations found in the memory are available in a pop-up window when required ; each is annotated with its status (for example, whether it is a standard equivalent term, has been approved by an in-house terminologist, has been accepted in a previous translation, etc.) Status may be changed, marking a given translation as standard or preferred, or even removing one from the memory if it is found to be unsuitable.

User-initiated searches with the translation memory are also provided for, allowing a translator to view all occurrences of a given expression in context.

When a translation is complete, newly translated segments are inserted into the translation memory ; since its inception, the CIPO application has created a translation memory of around 3.5 million segments.

2.5 Auxiliary Functions

2.5.1 Management

A project manager's interface supports assignment of tasks to available translators, and provides access to automatically tabulated translator performance statistics and progress data.

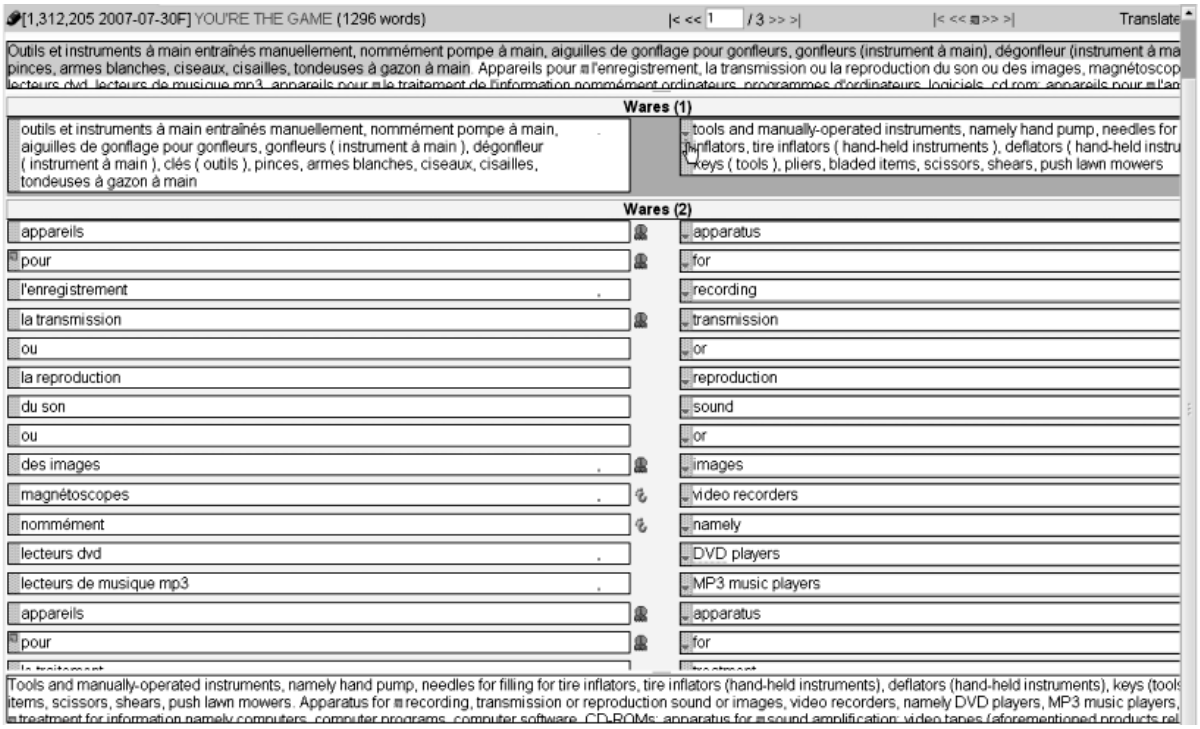

FIG. 1 – Moz translation editing interface immediately after pretranslation : source text to the left, target to the right, medallion icons indicating approved terminological equivalents. Original source text is shown above, and the current state of the target text below the main display.

2.5.2 Communications

During the translation phase, the translator may annotate a segment with comments or queries to be viewed by a revisor or other members of the translation team. Items needing special attention from a revisor may be optionally flagged. Similar annotations may also be created for a document as a whole.

A separate interface is available for off-line discussion of terminology questions by team members ; these are recorded for later reference.

In some circumstances it is necessary to communicate with the client in order to clarify or correct the input text ; a third dedicated communication facility is provided for this purpose.

Références

COUTURIER J.-F., NEUVEL S. & DROUIN P. (2006). Applying lexical constraints on morpho-syntactic patterns for the identification of conceptual-relational content in specialized texts. In *Proceedings of LREC 2006*, p. 591–594.