When to choose SMT

Typology of Documents

François Lanctôt, CEO
SilexCreations Inc.

BACKGROUND INFORMATION

SilexCreations is a provider of multimedia, multiplatform and multilingual services, founded in 1995

Firm carries out translation activities; also focuses on Audio-Visual and Web/Mobile applications. R&D (audio analytics)

User of **Rule-Based translation system** (Systran) since 1995

Part of a **Statistical Machine Translastion** pilot (PORTAGE) in 2009-2010

Licensee of **PORTAGE** commercial version since December 2013

WHAT IS PORTAGE?

Acronym stands for *Probabilistically Optimized Rules for Translation Automatically Generated from Examples*

Developed by the National Research Council of Canada (on-going project)

Main language pairs: English->French, French->English (official languages in Canada)

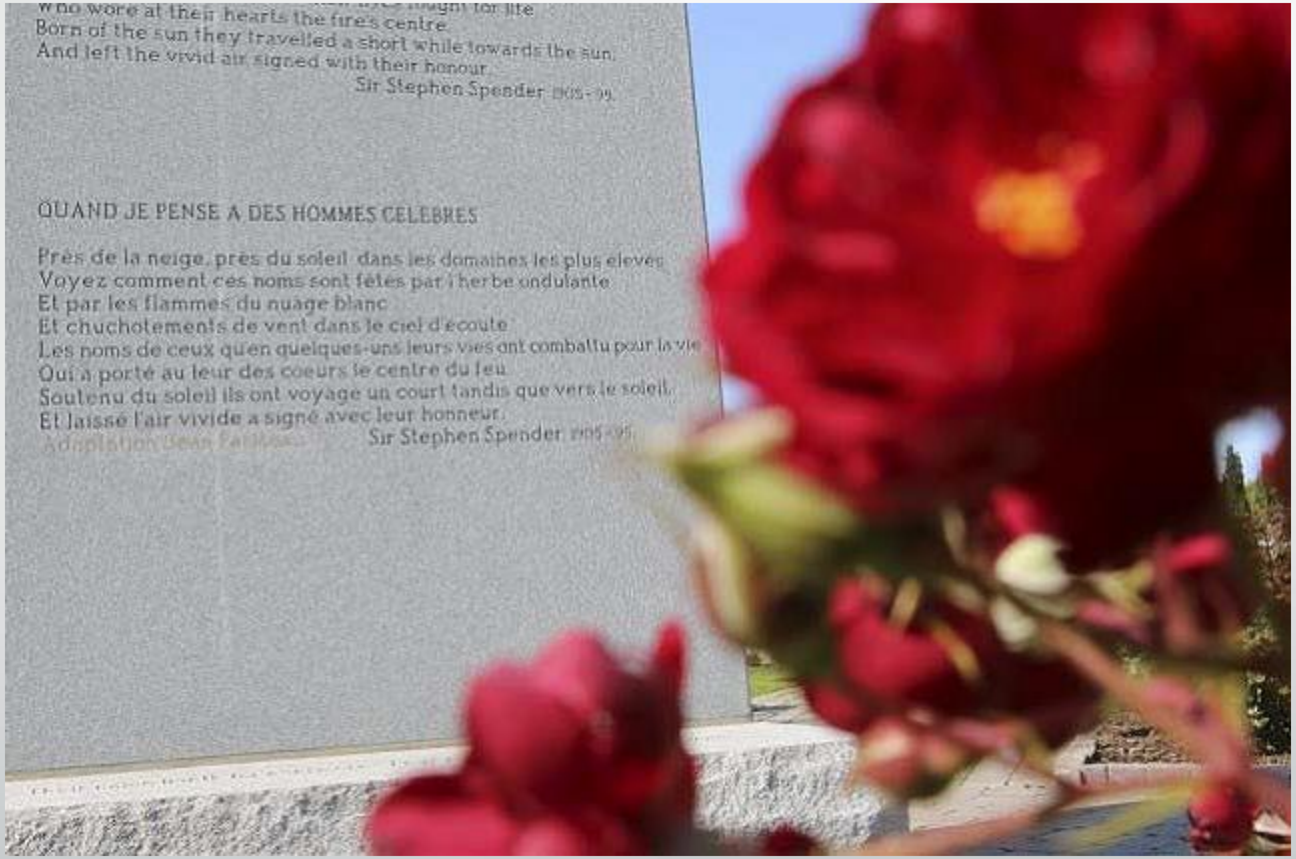
Other language pairs include: Mandarine->English, Danish->English, Arabic->English

No cross-linguistic capability

Open Machine Translation Evaluation (NIST): high BLEU scores in all categories

Typology of Documents

A phenomenal mistake: engraved poem went through Google Translate!!!



fought for life
Who wore at their hearts the fire's centre
Born of the sun they travelled a short while towards the sun,
And left the vivid air signed with their honour.
Sir Stephen Spender 1909 - 95

QUAND JE PENSE A DES HOMMES CELEBRES

Près de la neige, près du soleil dans les domaines les plus élevés
Voyez comment ces noms sont fêtés par l'herbe ondulante
Et par les flammes du nuage blanc
Et chuchotements de vent dans le ciel d'écoute
Les noms de ceux qu'en quelques-uns leurs vies ont combattu pour la vie
Qui a porté au leur des coeurs le centre du feu
Soutenu du soleil ils ont voyage un court tandis que vers le soleil,
Et laissé l'air vivide a signé avec leur honneur.
Adaptation Jean [illegible] Sir Stephen Spender 1909 - 95

BEST CASE SCENARIO

By default, PORTAGE is trained with the HANSARD bilingual corpus, i.e. transcripts of the Parliamentary Debates in the House of Commons (Canada)

Consequently, any sentence dealing with the terminology and phraseology specific to this corpus will yield an acceptable output requiring minimal post-editing.

E.g.: *The Hon. MP should understand that no privilege, whatsoever, is to be granted to the population of the province of Nova Scotia, nor to the Aboriginal people inhabiting the Northern parts of or our country, who remain for the most part in a situation of extreme indigence.*

With proper training, the system can even be configured for a speech-to-speech application. Europarl: Parallel corpus

STYLE and VOCABULARY

Conclusion: the writing style and vocabulary of the source documents have a huge impact on the end-result (and on post-editing).

As opposed to Google Translate, which uses a «one-size-fits-all» approach, Portage requires domain-specific corpora for training. The writing style and subject matter are critical.

Portage does NOT allow incremental training. This may be an advantage, because it avoids unwanted situations of ambiguity, which the system cannot resolve.

Training must always be performed in one batch, using the material with the appropriate style and vocabulary for a given translation project. As usual with SMT, addition of dictionaries or lexicons is not supported.

REQUIREMENTS FOR TRAINING

Bank of bilingual documents with vocabulary intensive sentences or segments.

According to Portage developers, the minimum size of a viable corpus is 200,000 words (source language only). Prepared with optimal coverage

A tool for text categorization may be helpful to determine the relevance and consistency of documents to be included in a corpus.

The time for aligning the source and target material within the training corpus should be taken into consideration, because it involves several adjustments (*.TMX format, PDF extraction and deformatting, etc.)

Portage is based on Linux CentOS and the command line syntax for training the system requires some expertise, and powerful hardware.

TRAINING: BASIC STEPS

Create language model

Create translation model

Optimizing decoder weights

Fine tuning

LIMITATIONS

Noun groups: even if 2 separate terms are present in the training corpus, there is no guarantee that the translation model will interpret them correctly when joined in a collocation.

Adjectives preceeding a term, e.g.: *this stupid and useless war* (outputs wrong sequence in French) VS *this stupid war* or *this stupid Holy war*: language model has a problem with unpredictable combinations of adjectives (a very frequent situation).

Single word or syntagma without context: in the absence of surrounding determinants, ambiguity cannot be resolved (e.g.: *Putting Rules*.

Impossible to 'force' a collocation or a noun group into the translation, unless it is pre-edited in the source text (using underscores) and replaced globally in the translated output.

IDIOMS and PHRASAL VERBS

Portage can handle idiomatic expressions, provided that they are included in the training corpus. Sentences like *Please bring me up to speed on this topic* or *He is totally out to lunch* will sound natural in French, better than with Google Translate or Bing. Whereas *All things must pass!* or *Let bygones be bygones*, unknown to the system, will create total nonsense in French (same with Bing).

Double and multiple meanings: a challenge for any SMT system. Frequent English verbs such as *to draw* or *to get* may be compounded with several prepositions, creating ambiguities.

Other verbs can have opposite significations, based on the context. E.g. *to clip.* A real mind twister for both Portage and Google Translate!

POST-EDITING

Best practice: post-editing with a TranslationMemory tool.

Extraction of untranslated segments in TMX format

Submission of segments to Portage's SMT

Reintegration of machine translated segments into translation memory

Filtering threshold for confidence (optional)

Dealing with levels of quality VS rewriting (human decisions)

SOAP interface for integration with any other tool or platform

PORTAGE IN CANADA

Translation Bureau of Public Works and Government Services Canada
http://www.bt-tb.tpsgc-pwgsc.gc.ca/btb.php?lang=eng

Major translation agencies

Confidentiality of training corpora and submitted documents