Asia Online : des outils de traduction très sophistiqués

ICIC 2014

Francois Libmann

On parle depuis plusieurs dizaines d'années de traduction automatisée. Les progrès ont été constants et de plus en plus de traductions machine de qualité satisfaisante deviennent disponibles. C'est le cas, par exemple pour les brevets.

Compte-tenu du développement de ce domaine, nombreux sont les acteurs qui proposent une offre.

L'une des entreprises très active sur le sujet est Asia Online basée à Singapour et très bien implantée en Asie dont le CEO Dion Wiggins, l'un des fondateurs de la société en décembre 2006, a fait une passionnante présentation, qu'il a orientée, public oblige, sur les brevets.

Pour bien positionner les choses, le conférencier a d'emblée précisé que leur outil "Language Studio" était une plateforme de traitement du langage et non pas un simple outil de traduction. Il peut être utilisé en SAAS ou implanté chez le client.

Quelques éléments dressent un premier portrait d'Asia Online :

- Asia Online est aujourd'hui capable de traiter 534 paires de langues.
- Le tout premier client a été LexisNexis Univentio en 2008 et le premier moteur a été vendu pour traduire en anglais des brevets japonais.
- L'implantation en Thaïlande a certainement contribué au projet lancé au tout début 2011 avec des partenaires publics et privés d'une traduction en langue thaï des 3,6 millions d'articles en anglais de l'encyclopédie Wikipedia.
- L'ensemble des utilisateurs de la plateforme traduit plus de 2 milliards de mots par jour, un seul de ces clients d'Asia Online en traduisant plus de la moitié dans un corpus brevet.
- La plus grande vitesse de traitement demandée par un client (en l'occurrence un gouvernement) était de 600 millions de mots par minute.
- On notera aussi que l'industrie du traitement du langage ne traduit que 0,00000067% des textes créés chaque jour dans le monde.
- Un traducteur humain mettrait 152 257 ans à traduire en anglais tous les brevets japonais existants et cela coûterait 40 Milliards de dollars.
- La phrase la plus longue trouvée dans un brevet compte 4 500 mots.

L'un des grands principes d'Asia Online est que l'on ne peut faire de la qualité dans ce domaine qu'à la condition de comprendre les données. Cela explique en le recours particulier, à deux catégories d'experts, ceux du domaine traité et ceux des langues concernées. Plus généralement l'humain intervient à plusieurs niveaux du processus pour garantir la qualité.

D'autre part, pour alimenter son SMT (Statistical Machine Translation), Asia Online revendique le choix d'un modèle SMT avec des données « propres » plutôt qu'avec des données « sales » ce qui veut dire :

- que les données proviennent d'un petit nombre de sources dont on peut faire confiance dans la qualité plutôt que de chercher à réunir un maximum de sources,
- que les données doivent provenir du même domaine que celui dans lequel on va faire la traduction
- que la qualité des données est plus importante que leur quantité.

Le conférencier a d'ailleurs précisé que si tous les clients d'Asia Online ne sont pas dans le domaine des brevets, c'est certainement le contenu le plus complexe à traiter.

A titre d'exemple, pour analyser un brevet, les codes IPC permettent de définir à quelle catégorie, parmi cinq, un brevet appartient, les styles des textes étant propres à chaque domaine de connaissance.

Par ailleurs, les styles utilisés dans le titre, l'abstract, les revendications et la description sont également différents.

On arrive donc à 20 combinaisons possibles pour pouvoir utiliser le « sous-moteur » le mieux adapté à la fois au domaine et à la partie du brevet que l'on veut traduire.

Nous n'entrerons pas dans les détails et les différents cas nécessaires au "nettoyage" des données.

Une bonne illustration de l'importance du contexte pour le style peut être donnée dans l'exemple suivant où l'on part d'un document original en espagnol pour arriver à sa traduction en anglais dans deux contextes très différents. Le premier est un contexte business illustré par The Economist, le New York Times ou Forbes et le second celui d'un livre pour enfant.

Original en espagnol :

Se necesito una gran maniobra politica muy prudente a fin de facilitar una cita de los dos enemigos historicos.

Traduction dans un contexte business :

Significant amounts of cautious political maneuvering were required in order to facilitate a rendezvous between the two bitter historical opponents.

Traduction dans un contexte de livre pour enfants :

A lot of care was taken to not upset others when organizing the meeting between the two long time enemies.

On notera que Questel utilise Asia Online parmi d'autres fournisseurs de traduction machine.