Stage MR2 LID 2007-2008

Encadrant :
Yves Lepage
GREYC, CNRS UMR 6072, bureau S3-383,
tél : 02 31 56 74 82

Réimplémentation et amélioration d'un système de traduction automatique par analogie

---

Contexte :

La traduction automatique du tout-venant sur le Web est déjà une réalité avec l'usage de Systran (site propre, BabelFish, etc.). Systran est un système fondé sur les connaissances linguistiques a priori. Son développement continue jusqu'à nos jours et il a requis plusieurs centaines d'hommes x années depuis plus de cinquante ans. Le développement d'un tel système est donc extrêmement lourd.

Une nouvelle approche de la traduction automatique consiste à partir de données bilingues pour construire des systèmes plus légers, et dédiés à un domaine particulier (par exemple, le domaine du tourisme). On dit de tels systèmes qu'ils sont fondés sur les données. L'approche la plus en vogue à l'heure actuelle est l'approche statistique avec les modèles d'IBM et dont on peut voir la mise en application avec Google Traduction. Une autre approche fondée sur les données est l'approche à base d'exemples. Le système de traduction automatique ALEPH développé à ATR au Japon puis à l'université de Caen est un tel système. Il a comme principe de base la correspondance entre analogies. Pour traduire, il faut mettre la phrase à traduire en relation d'analogie avec d'autres phrases de la même langue, et, si on connaît les traductions de ces phrases, résoudre une équation analogique dans la langue d'arrivée (voir plus bas).

Pour mesurer les performances des systèmes basés sur les données, des campagnes d'évaluation sont organisées régulièrement. Le système ALEPH a participé aux campagnes d'évaluation IWSLT 2005 et IWSLT 2007.

Objectif :

Le ou les étudiants devront suivre un objectif très pratique : réimplémenter le système ALEPH dans un langage de plus haut niveau, Python, en adoptant des solutions de génie logiciel.

Cette réécriture inclut deux volets de recherche ardus :

- la réimplémentation de l'opération de base, l'analogie, selon une nouvelle formalisation. L'étude de cette nouvelle formalisation pourra s'accompagner de la résolution d'une question théorique importante : les langages de parenthèses sont-ils des langages analogiques ?
- la réduction de l'espace de recherche lors de la traduction. Des techniques de fouille de données pourront être étudiées. En plus, le stage devra déboucher sur la mise en place de techniques de visualisation de l'espace de recherche par matrices analogiques.

Environnement :

Ce stage demandera un investissement fort. Il faut le considérer comme le début d'un travail de thèse sur le sujet de l'alignement-traduction et de la traduction automatique.

Le but est que le système réimplémenté soit mis à la disposition de la communauté de recherche en traduction automatique. De plus, le nouveau système participera à la campagne d'évaluation IWSLT 2008 qui rassemble les équipes de recherche les plus fortes du domaine. On s'y mesure essentiellement à des laboratoires américains ou japonais et à Systran. Ce stage est donc une occasion rêvée, pour un ou des étudiants ambitieux, d'entrer en contact le plus rapidement possible avec les meilleurs laboratoires du domaine. Si les résultats sont bons, le stage conduira à la publication d'articles scientifiques, ce qui entre en compte dans les critères d'attribution des bourses de doctorat.

Bibliographie

- Etude du système ALEPH : (Lepage et Denoual, 2005) décrit le système.

- Etude de l'analogie entre chaînes de symboles : (Lepage, 2003) donne les fondements historiques et théoriques sur l'analogie proportionnelle entre chaînes de symboles.

- Etude des campagnes d'évaluation : voir les sites d'IWSLT 2005 et d'IWSLT 2006 ; (Lepage et Denoual, 2005) donne les résultats du système ALEPH sur les données des campagnes IWSLT 2005 et IWSLT 2006.

- Etude des matrices analogiques : (Lepage et Peralta, 2003) décrit un travail de production de matrices analogiques dans la perspective de la génération de paraphrases.

Précision sur l'analogie :

Une analogie met quatre phrases en relation : la première phrase est à la seconde ce que la troisième phrase est à la quatrième. Par exemple en anglais :

Could you cash a traveler's check? : I'd like to cash these traveler's checks. :: Could you open a window? : I'd like to open these windows.

ou bien en français :

vous pouvez m'échanger un chèque de voyage ? : ces chèques de voyage, là, je peux les échanger ? :: est-ce que vous pouvez m'ouvrir un volet ? : est-ce que ces volets, là, je peux les ouvrir ?

La correspondance s'établit entre quatre phrases de la langue de départ en relation d'analogie et quatre phrases de la langue d'arrivée aussi en relation d'analogie. Par exemple, avec les phrases ci-dessus :

Could you cash a traveler's check? <----> *vous pouvez m'échanger un chèque de voyage ?*

I'd like to cash these traveler's checks. <----> *ces chèques de voyage, là, je peux les échanger ?*

Could you open a window? <----> *est-ce que vous pouvez m'ouvrir un volet ?*

I'd like to open these windows. <----> *est-ce que ces volets, là, je peux les ouvrir ?*

Précision sur la méthode de traduction par analogie :

Pour traduire, supposons que l'on connaisse les trois correspondances de traduction suivantes :

I'd like to cash these traveler's checks. <----> *ces chèques de voyage, là, je peux les échanger ?*

Could you open a window? <----> *est-ce que vous pouvez m'ouvrir un volet ?*

I'd like to open these windows. <----> *est-ce que ces volets, là, je peux les ouvrir ?*

Soit la phrase à traduire *Could you cash a traveler's check?*. Sa traduction est inconnue : *x*. On peut la placer dans la relation d'analogie suivante :

Could you cash a traveler's check? : *I'd like to cash these traveler's checks.* :: *Could you open a window?* : *I'd like to open these windows.*

En prenant les phrase en correspondance de traduction, on forme l'équation analogique suivante d'inconnue *x* :

x : *ces chèques de voyage, là, je peux les échanger ?* :: *est-ce que vous pouvez m'ouvrir un volet ?* : *est-ce que ces volets, là, je peux les ouvrir ?*

La résolution de cette équation donne :

x = vous pouvez m'échanger un chèque de voyage ?

Par hypothèse de correspondance de traduction entre analogies dont les termes sont en correspondance de traduction, on conclut que la traduction de la phrase *Could you cash a traveler's check?* est *vous pouvez m'échanger un chèque de voyage ?*