Sujet de stage ENS Cachan

Encadrant : Yves Lepage, S3-383
yves.lepage@info.unicaen.fr

**Structuration d'ensembles de chaînes de symboles :
réduction de l'espace de recherche pour la traduction automatique par analogie**

Contexte :

Il existe deux approches en traduction automatique : l'approche par règles et l'approche fondée sur les données. L'approche par règles est illustrée par Systran, système commercial en situation de quasi-monopole (Yahoo!, AltaVista Babelfish, etc.). Dans l'approche fondée sur les données, on distingue l'approche probabiliste mise en oeuvre par Google avec ses systèmes disponibles depuis mi-novembre 2007 et l'approche par l'exemple. Ce stage a pour but d'améliorer un système de traduction par l'exemple développé au GREYC, le système ALEPH.

Aussi bien l'approche probabiliste que l'approche par l'exemple recourent à d'importantes quantités de données d'entraînement. Le système ALEPH participe régulièrement à la campagne d'évaluation de la traduction automatique IWSLT au cours de laquelle sont mises à disposition 40 000 phrases courtes dans le domaine du tourisme dans plusieurs langues.

Le système ALEPH repose essentiellement sur l'opération d'analogie qui met en relation quatre phrases, c'est-à-dire quatre chaînes de caractères, comme dans l'exemple suivant. La formalisation (partielle) de l'analogie utilisée ici fait essentiellement appel aux distances d'édition et au comptage de caractères. En une seconde, on peut vérifier plusieurs centaines de milliers d'analogies.

Une tasse de café, s'il vous plaît. : Deux cafés, s'il vous plaît. :: Une tasse de thé au lait. : Deux thés au lait.

Pour traduire une phrase nouvelle, le principe de base du système ALEPH est de mettre cette phrase en relation d'analogie avec trois phrases des données d'entraînement. Le problème serait donc cubique en la taille des données d'entraînement (taille de l'espace de recherche : 40 000^3).

Heureusement, l'opération d'analogie est aussi productive, c'est-à-dire que trois phrases quelconques (ou chaînes de caractères) forment une équation analogique que l'on peut toujours tenter de résoudre. Selon les cas, on a zéro, une ou plusieurs solutions. Chaque solution est une chaîne de caractères (donc éventuellement une phrase). Pour traduire une phrase nouvelle, on peut donc former toutes les équations analogiques possibles avec la phrase à traduire et tous les couples possibles construits sur les données d'entraînement (nouvelle taille de l'espace de recherche : 40 000^2 = 1,6 milliard). Il « suffirait » alors de vérifier si les solutions des 1,6 milliards d'équations analogiques appartiennent aux données d'entraînement ou pas.

Un milliard et demie de couples est encore de l'ordre de l'impraticable. On propose donc de structurer l'espace de recherche en groupant les phrases par échelles d'opposition, comme illustré ci-dessous.

Un café, s'il vous plaît. : Un café.
Un thé, s'il vous plaît. : Un thé.
Apportez-moi le journal, s'il vous plaît. : Apportez-moi le journal.
Quelle heure est-il, s'il vous plaît ? : Quelle heure est-il ?

Ces échelles d'opposition sont en nombre raisonnable (quelques centaines seulement pour plusieurs milliers de phrases). Il est donc envisageable d'échantillonner ces échelles pour n'avoir plus à explorer qu'un nombre praticable d'équations analogiques.

But du stage :

On proposera des algorithmes pour la production des échelles de phrases en opposition, on implémentera ces algorithmes et on les appliquera à des données issues de campagnes d'évaluation de systèmes de traduction automatique (anglais, japonais, arabe). On échantillonnera alors les échelles pour modifier le système ALEPH et l'accélérer. On mesurera les performances du nouveau système obtenu, en qualité et en vitesse, sur des données de test et on comparera avec l'ancien système.

On explorera aussi la possibilité d'encoder chaque phrase des données d'entraînement selon sa présence dans les échelles d'opposition à l'aide d'un code binaire. Placer une phrase nouvelle dans l'ensemble des données d'entraînement, c'est lui attribuer un code sur toutes les dimensions correspondant aux différentes échelles.

Bibliographie :

Langlais, P. and Patry, A., Translating Unknown Words by Analogical Learning, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 877-886.

Lepage, Y., and Denoual, E. Purest ever example-based machine translation: detailed presentation and assessment. Machine Translation Journal (2005), p. 251-282.

Stroppa, N. and Yvon, F., Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie, Traitement Automatique des Langues, 47(2):33-59, 2006.