iMAG : post-édition, évaluation de qualité de TA et production d'un corpus parallèle

Lingxiao WANG Ying ZHANG
GETALP – LIG, 41 rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9
Lingxiao.Wang@imag.fr, Ying.Zhang@imag.fr

RÉSUMÉ

Une passerelle interactive d'accès multilingue (iMAG) dédiée à un site Web S (iMAG-S) est un bon outil pour rendre S accessible dans beaucoup de langues, immédiatement et sans responsabilité éditoriale. Les visiteurs de S ainsi que des post-éditeurs et des modérateurs payés ou non contribuent à l'amélioration continue et incrémentale des segments textuels les plus importants, et éventuellement de tous. Dans cette approche, les pré-traductions sont produites par un ou plusieurs systèmes de Traduction Automatique (TA) gratuits. Il y a deux effets de bord intéressants, obtenables sans coût additionnel : les iMAGs peuvent être utilisées pour produire des corpus parallèles de haute qualité, et pour mettre en place une évaluation permanente et finalisée de multiples systèmes de TA.

ABSTRACT

iMAG : MT-postediting, translation quality evaluation and parallel corpus production

An interactive Multilingual Access Gateway (iMAG) dedicated to a web site S (iMAG-S) is a good tool to make S accessible in many languages immediately and without editorial responsibility. Visitors of S as well as paid or unpaid post-editors and moderators contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. In this approach, pre-translations are produced by one or more free machine translation systems. There are two interesting side effects obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora and to set up a permanent task-based evaluation of multiple MT systems.

MOTS-CLÉS : post-édition, évaluation de systèmes de TA, production d'un corpus parallèle
KEYWORDS : post-edition, evaluation of MT systems, production of parallel corpora.

Nous proposons 3 démonstrations : (1) l'accès multilingue à un site Web, avec la post-édition de résultats de TA "à la Google" ; (2) la post-édition en mode avancé; (3) la production d'un corpus parallèle.

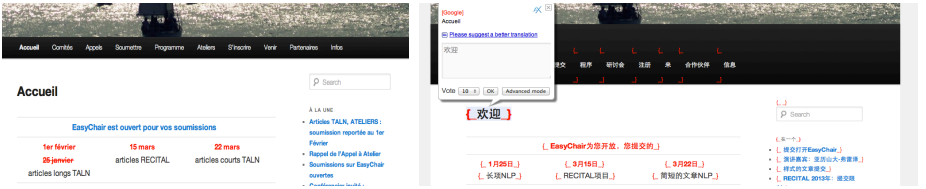

FIGURE 1 – Page originale en anglais et page accédée en chinois.

Voici un exemple d'accès au site Web de TALN 2013 en chinois. L'original est en anglais

comme montré en figure 1. Nous choisissons le chinois dans le menu déroulant et cochons la case "Reliability". La page est désormais accessible en chinois, avec des parenthèses spéciales autour des segments. Les traductions initiales sont réalisées par un ou plusieurs serveurs de TA gratuits. Dans ce cas, nous utilisons Google Translate et Systran. Lorsque le curseur passe sur un segment, un tableau apparaît, à travers lequel la post-édition peut être effectuée directement, "sans couture". Le mode avancé de PE consiste à post-éditer un pseudo-document qui est en fait une partie de la mémoire de traductions (MT).

Dans la figure 2, le premier segment a été pré-traduit par Google Translate, et le deuxième segment a été post-édité. Nous pouvons voir la MT (pré-traductions et post-éditions), et voir la « distance d'édition » pour chaque segment, entre chaque pré-traduction ou post-édition différente alternative et le texte source.

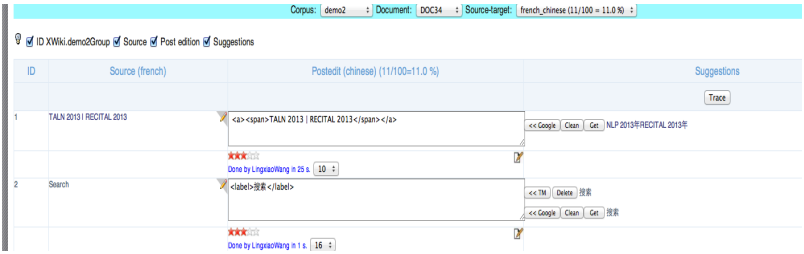

FIGURE 2 – Post-édition en mode avancé (capture d'écran de SECTra_w).

Grâce à SECTra_w, qui offre un système d'annotation de chaque traduction ou post-édition d'un segment par un niveau de fiabilité (de * à *****) et un score de qualité (de 0 à 20), il est possible d'extraire de la mémoire de traductions, associée à un site Web S, une sous-MT vérifiant n'importe quel prédicat basé sur les niveaux et les scores.

L'exemple suivant (figure 3) montre une extraction simple, à partir de la partie français-chinois de la MT-Demo2. Le prédicat est [Level = 3 & score> = 13], et ses paramètres peuvent être choisis directement via l'interface graphique. La sélection peut être exportée (comme montré en figure 4), en 2 fichiers parallèles, dans un format XML simple, utilisée plus tard comme corpus supplémentaire d'apprentissage d'un système de TA empirique (comme Moses-LIG) pour être spécialisé à ce site Web.

FIGURE 3 – Extraction d'une « bonne » MT de la MT produite par post-édition « naturelle »

FIGURE 4 – Export d'une « bonne » MT

Références

HUYNH, C.-P., BOITET, C., BLANCHON, H. & NGUYEN, H.-T. (2009). SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08*, Marrakech, 27-31/5/08, ELRA/ELDA, ed., 8 p.