ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs

Francis Brunet-Manquat[1] Emmanuelle Esperança-Rodier[1]
(1) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

Francis.Brunet-Manquat@univ-grenoble-alpes.fr,Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

RESUME

La plateforme ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs) propose une palette de services innovants permettant de répondre aux besoins modernes d'analyse d'erreurs de traduction : gestion simplifiée des corpus et des typologies d'erreurs, annotation d'erreurs efficace, collaboration et/ou supervision lors de l'annotation, recherche de modèle d'erreurs dans les annotations.

ABSTRACT

ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpus. We present a platform for the collaborative editing of translation errors. This platform, named ACCOLÉ, offers a range of innovative services that meet the analysis needs of translation errors: simplified management of corpora and typologies of errors, annotation of effective errors, collaboration and/or supervision during annotation, looking for error types in annotations.

MOTS-CLES : Annotations d'erreurs de traduction, Annotation collaborative
KEYWORDS: Annotations of translation errors, Collaborative annotation

La plateforme ACCOLÉ permet l'annotation manuelle des erreurs de traduction selon des critères linguistiques basés sur la typologie de (Vilar et al., 2006). Elle se situe dans la lignée des travaux portant sur l'estimation de la qualité comme QuEst++ (Specia et al., 2015) et de l'analyse d'erreurs tels que Coreference Annotator (Tsoumari et al., 2001) ou BLAST (Stymne, 2011). Les travaux de (Esperança et al., 2016) montrent les limites de ce dernier quant à son utilisation pour la tâche que nous nous sommes fixée. Il s'agit de fournir à un utilisateur une aide dans le choix d'un système de TA à utiliser selon le contexte (compétences linguistiques et informatiques de l'utilisateur, connaissance du domaine du document source à traduire et la tâche pour laquelle il a besoin de traduire le document source.). ACCOLÉ doit donc permettre de détecter quels sont les phénomènes linguistiques qui ne sont pas traités correctement par le système de TA étudié. Nous proposons ainsi, sur la même plateforme une palette de services innovants permettant de répondre aux besoins modernes d'analyse d'erreurs de traduction. Ainsi, les principales fonctionnalités de la plateforme ACCOLÉ sont la gestion simplifiée des corpus, des typologies d'erreurs, des annotateurs, etc. ; l'annotation d'erreurs efficace ; la collaboration et/ou supervision lors de l'annotation ; la recherche de modèle d'erreurs (type d'erreurs dans un premier temps) dans les annotations. La plateforme est disponible en ligne sur un navigateur et ne nécessite aucune installation spécifique.

1 Annotation d'erreurs

La plateforme ACCOLÉ propose de visualiser et d'annoter efficacement les erreurs d'un couple de phrases source/cible. L'annotation se fait en deux étapes. La première étape consiste à sélectionner, à l'aide de la souris, des mots dans la phrase source, et de leur équivalent dans la phrase cible, présentant une erreur de traduction. La seconde étape consiste à choisir le type d'erreur à associer au couple des mots sources/cibles préalablement sélectionnés. En plus de sa simplicité d'usage, ACCOLÉ propose deux mécanismes pour aider l'annotateur dans sa tâche : un mécanisme de supervision permettant à un responsable de contrôler l'avancée de la tâche, ce mécanisme encourage surtout la communication entre superviseur et annotateur par la possibilité de créer des fils de discussion pour un couple de phrases source/cible précis (demander des précisions sur un type d'erreur, pointer une erreur d'annotation, etc.); et un mécanisme collaboratif permettant aux annotateurs de s'entraider ou de discuter autour d'un couple phrase source/cible précis.

2 Représentation des erreurs basée sur les SSTC

La plateforme utilise une représentation des données basée sur les SSTC (Structured String-Tree Correspondences, Boitet et Zaharin 1988). Une erreur est donc constituée d'une étiquette et d'un ensemble de SNODE (intervalle représentant la sous-chaîne dans la phrase source ou cible correspondante). Par exemple dans la figure ci-dessous représentant un exemple d'annotations, l'erreur portant sur "toute l'" et "any" est décrite par son étiquette *Mauvais choix lexical* (cat. *Mot incorrect*, sous-cat. *Sens*), par son positionnement dans la phrase source (SNODE [49-56] - sous chaîne entre le 49ème caractères et le 56ème) et la phrase cible (SNODE [46-48]). L'un des avantages d'utiliser ainsi les SNODE est de se passer d'une structure syntaxique pour décrire l'erreur.

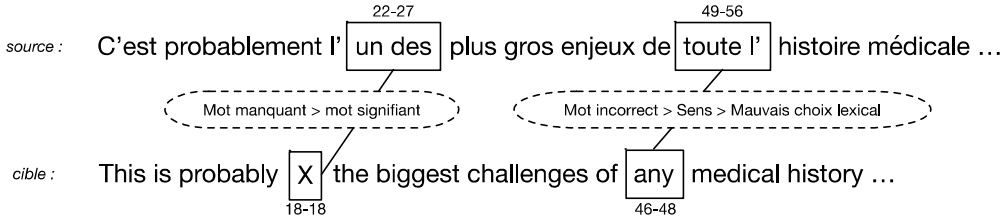

FIGURE 1: exemple d'annotations

3 Recherche d'erreurs

La recherche de type d'erreurs dans les annotations est un atout essentiel de la plateforme. Nous souhaitons l'améliorer en permettant, en plus de l'association d'analyses morphosyntaxiques aux annotations d'erreurs, l'association d'arbres de dépendances, ainsi que l'utilisation d'autres typologies d'erreurs telles que décrites dans Felice (2012) ou bien des métriques multidimensionnelles de qualité (MQM) du projet QT21 (QT21, 2016).

Références

Boitet, C., Zaharin, Y. (1988). Representation trees and string- tree correspondences. *Proc. COLING-88*, 59-64.

Esperança-Rodier E., Didier, J. 2016. Translation Quality Evaluation of MWE from French into English using an SMT system. *Actes de la 38th Conference Translating and the Computer, London, UK, November 17-18, ©2016 AsLing 33–41*

Felice M., Specia L. (2012). Linguistic Features for Quality Estimation. Actes de *7th Workshop on Statistical Machine Translation*, Montréal, Canada, June 7-8, 2012 Association for Computational Linguistics, 96–103

QT21: Quality Translation 21. Available at: http://www.qt21.eu/. Accessed: September 25, 2016.

Specia L., Paetzold G. H. et Scarton C. (2015): Multi-level Translation Quality Prediction with QuEst++. Actes de *ACL-IJCNLP 2015 System Demonstrations*, Beijing, China, 115-120.

Stymne S.(2011). Blast: A Tool for Error Analysis of Machine Translation Output. Actes de *ACL-HLT 2011 System Demonstrations*. June 19-24, 2011. Portland, Oregon, USA, 56-61

Tsoumari, M. and Petasis, G. 2001. Coreference Annotator - A new annotation tool for aligned bilingual corpora. Actes du *Second Workshop on Annotation and Exploitation of Parallel Corpora* (AEPC 2), dans 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)

Vilar D., Xu J., D'Haro L. F., Ney H. (2006). Error Analysis of Statistical Machine Translation Output. Actes *LREC*, Genoa, Italy, 697-702