Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Contrôle de la Terminologie en Traduction Automatique Neuronale

Melissa Ailem Jingshu Liu Raheel Qader
Lingua Custodia, France
{melissa.ailem,jingshu.liu,raheel.qader}@linguacustodia.com

RÉSUMÉ

Nous présentons une nouvelle approche permettant d'intégrer des contraintes terminologiques dans les modèles de traduction neuronale. Notre méthode agit pendant la phase d'entraînement évitant ainsi toute augmentation du temps de calcul pendant l'inférence. L'approche proposée combine trois ingrédients essentiels. Le premier consiste à augmenter les données d'apprentissage afin de spécifier les contraintes. Intuitivement, cela permet au modèle d'apprendre à copier les contraintes dans la traduction prédite. Contrairement aux travaux existants, nous utilisons une technique d'augmentation de données simplifiée, ne nécessitant pas l'utilisation de *source factors*. Le second ingrédient consiste à masquer le terme source des contraintes, permettant au modèle d'apprendre encore plus facilement le comportement de copie et de mieux se généraliser. Le troisième est une modification de la fonction standard d'entropie croisée afin d'encourager le modèle à attribuer des probabilités élevées aux mots appartenant aux contraintes. Les résultats montrent l'efficacité de notre approche en termes de score BLEU et en termes de pourcentage de contraintes respectées.

ABSTRACT

Encouraging Neural Machine Translation to Satisfy Terminology Constraints.

We present a new approach to encourage neural machine translation to satisfy lexical constraints. Our method acts at the training step and thereby avoiding the introduction of any extra computational overhead at inference step. The proposed method combines three main ingredients. The first one consists in augmenting the training data to specify the constraints. Intuitively, this encourages the model to learn a copy behavior when it encounters constraint terms. Compared to previous work, we use a simplified augmentation strategy without source factors. The second ingredient is constraint token masking, which makes it even easier for the model to learn the copy behavior and generalize better. The third one, is a modification of the standard cross entropy loss to bias the model towards assigning high probabilities to constraint words. Empirical results show that our method improves upon related baselines in terms of both BLEU score and the percentage of generated constraint terms.

MOTS-CLÉS : Traduction Automatique Neuronale, Contraintes Terminologiques, Augmentation des données d'apprentissage, score BLEU.

KEYWORDS: Neural Machine Translation, Lexical Constraints, Training Data Augmentation, BLEU score.

Volume 1 : conférence principale, pages 446–446.