Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Vers un contrôle plus robuste de la longueur de génération pour les modèles de langue auto-régressifs

Ivanhoé Botcazou Tassadit Amghar Sylvain Lamprier Frédéric Saubion
LERIA, UFR Sciences, 2 Bd de Lavoisier, 49000 Angers
{ivanhoe.botcazou, tassadit.amghar, sylvain.lamprier,
frederic.saubion}@univ-angers.fr

RÉSUMÉ

Les modèles de langue atteignent aujourd'hui un niveau remarquable en génération textuelle, cependant le contrôle précis de la longueur produite demeure un défi ouvert. Dans un premier temps nous revisitons la méthode RPE, fondée sur un décompte discret de positions inversées, informant le modèle du nombre de tokens lui restant à générer à chaque instant. Certaines instabilités liées à cette méthode et sa faible généralisation hors distribution, nous ont poussées à introduire un nouveau procédé nommé PRE. Notre méthode repose sur une représentation continue de l'avancement génératif en établissant un lien explicite entre des vecteurs d'« *impatience* » et des principes issus de la théorie du signal. PRE s'implémente via une modification légère de l'entrée d'un Transformer et offre un contrôle précis de la longueur cible, tout en préservant la qualité sémantique. Nos résultats expérimentaux s'appuient sur des tâches de résumé et de génération de questions.

ABSTRACT

Progress Ratio Embeddings : An Impatience Signal for Robust Length Control in Neural Text Generation

Modern neural language models achieve high accuracy in text generation, yet precise control over generation length remains underdeveloped. In this paper, we first investigate a recent length control method based on *Reverse Positional Embeddings (RPE)* and show its limits when control is requested beyond the training distribution. In particular, using a discrete countdown signal tied to the absolute remaining token count leads to instability. To provide robust length control, we introduce *Progress Ratio Embeddings (PRE)*, as continuous embeddings tied to a trigonometric *impatience* signal. PRE integrates seamlessly into standard Transformer architectures, providing stable length fidelity without degrading text accuracy under standard evaluation metrics. We further show that PRE generalizes well to unseen target lengths. Experiments on two widely used news-summarization benchmarks and a popular question generation dataset validate these findings.

MOTS-CLÉS : contrôle de longueur, modèles de langue, Transformer, résumé automatique, génération de questions.

KEYWORDS: length control, language models, Transformer, abstractive summarization, question generation.

ARTICLE ACCEPTÉ À : The 63rd Annual Meeting of the Association for Computational Linguistics.
URL : https://openreview.net/forum?id=s8M8eZM5f7

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 59–59.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.