

Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision

Wanyu Du, Zae Myung Kim, Vipul Raheja , Dhruv Kumar , Dongyeop Kang



— In2Writing

Workshop on Intelligent and Interactive
Writing Assistants (Best paper)

Introduction

La tâche: Text revision

Définition: « *text revision involves identifying discrepancies between intended and instantiated text, deciding what edits to make, and how to make those desired edits* »

- La révision d'un texte même pour les humain est **un processus itératif**
- Les modèles actuels sont des modèles *Seq2Seq one-shot "original-to-final"* -> **pas de modèle itératif**

Ce que propose le papier:

- ***R3 (Read, Revise, and Repeat)*** un modèle *human in the loop* pour faciliter et rendre plus efficace le processus de révision
- Des expériences pour **évaluer l'efficacité de *R3***
- Fournissent les **données issues des interactions humain/système** + leur **code** + leur **interface** pour de futures recherches

Approche: Modèle $\mathcal{R}3$

- **Étape 1:** Choix par l'utilisateur du document d'entrée
- **Étape 2:** Le système de révision de texte fournit des suggestions de modifications

Une suggestion= La modification + l'intention

- **Étape 3:** L'utilisateur accepte ou rejette les suggestions
- **Étape 4:** Fin de la révision si il n'y a plus de suggestions où que le nombre limite d'itérations est atteint. **Sinon retour à l'étape 2**

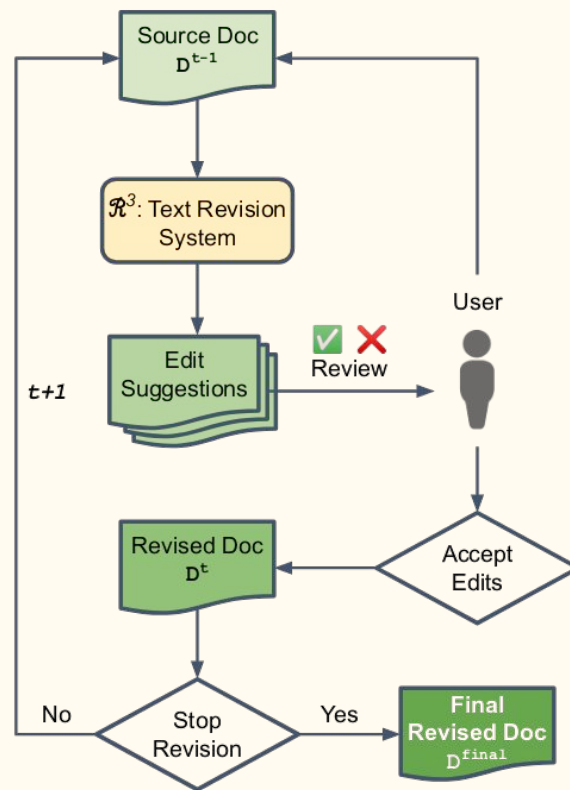


Figure 1: System overview for $\mathcal{R}3$ human-in-the-loop iterative text revision.

Les 3 modèles

Données utilisées pour le
fine-tuning des modèles

Pour chaque phrase 2 étapes:

- **Modèle n°1: edit-prediction classifieur**
 - **Faut-il modifier la phrase?** Oui ou Non (étiquette binaire)
- **Modèle n°2: edit-intention classifieur**
 - Si la phrase doit-être modifiée, **quelle intention appliquer?**
 - 4 intentions de modification: **FLUENCY**, **COHERENCE**, **CLARITY** et **STYLE**.

Modèles utilisés: RoBERTa-large fine-tuné

Modèle n°3: Text Revision Generation Model

- **Modèle utilisé:** PEGASUS fine-tuné
- **Entrée:** une phrase source et l'intention prédite
- **Sortie:** phrase révisée conditionnée par l'intention

	# Docs	Avg. Depths	# Edits
Training	44,270	6.63	292,929
Validation	5,152	6.60	34,026
Test	6,226	6.34	39,511

Table 1: Statistics for our collected revision data which has been used to train the edit intention identification model and the text revision generation model. **# Docs** means the total number of unique documents, **Avg. Depths** indicates the average revision depth per document (for the human-generated training data), and **# Edits** stands for the total number of edits (sentence pairs) across the corpus.

Exemples de modifications proposées par $\mathcal{R}_3$

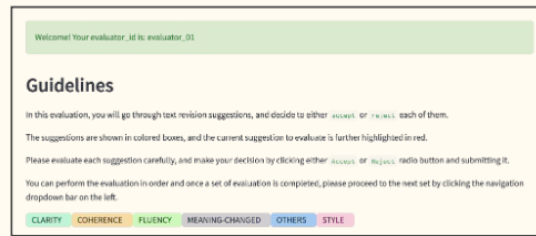
Edit Intention	Edit Suggestion
CLARITY	Emerging new test procedures such as antigen or RT-LAMP tests ; might enable us to protect nursing home residents.
FLUENCY	For Radar tracking, we show how a model can reduce the tracking errors.
COHERENCE	However, we show that even a small violation can significantly modify the effective noise.
STYLE	There has been numerous extensive research focusing on neural coding.

Table 4: Edit suggestion examples generated by $\mathcal{R}_3$. 4

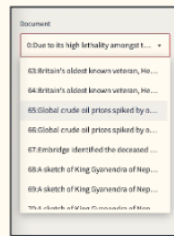
Interface utilisateur



(a) Login



(b) Read guidelines



(c) Select doc



(d) Editing suggestions and interaction panel

Figure 2: User interface demonstration for $\mathcal{R}3$. Anonymized version available at <https://youtu.be/1K08tIpEoaE>.

Expériences 1/3

Mesure d'évaluation: performance estimée en calculant le taux d'acceptation moyen des suggestions

- Les **3 questions** posées pour tester leur système:
 - **Q1: Est-ce que les modifications proposées sont de bonne qualité?** (=> Les utilisateurs ont-ils de fortes chances de les accepter?)
 - **Q2: Quels types de modifications ont le plus de chances d'être acceptées par les utilisateurs?**
 - **Q3: Est-ce efficace d'ajouter l'humain dans la boucle?** (=> Est-ce que R3 produit des modifications de meilleure qualité?)
- Pour y répondre **3 situations** sont comparées:
 - **Human-Human**
 - **System-Human**
 - **System-only**

Expériences 2/3

Comparaison des modifications proposées par un humain et par le système

Q1:

<i>t</i>	HUMAN-HUMAN				SYSTEM-HUMAN (ours)			
	# Docs	Avg. Edits	Avg. Accepts	% Accepts	# Docs	Avg. Edits	Avg. Accepts	% Accepts
1	30	5.37	2.77	51.66	30	5.90	2.90	49.15
2	30	4.83	3.00	62.06	24	3.83	2.57	67.02
3	20	3.80	2.67	70.39	20	3.43	1.94	56.71

Table 2: Human-in-the-loop iterative text revision evaluation results. *t* stands for the revision depth, # **Docs** shows the total number of revised documents at the current revision depth, **Avg. Edits** indicates the average number of applied edits per document, **Avg. Accepts** means the average number of edits accepted by users per document, and **% Accepts** is calculated by dividing the total accepted edits with the total applied edits.

Q2:

	HUMAN-HUMAN			SYSTEM-HUMAN (ours)		
	# Edits	# Accepts	% Accepts	# Edits	# Accepts	% Accepts
CLARITY	197	119	60.40	332	195	58.73
FLUENCY	178	146	82.02	91	41	45.05
COHERENCE	103	41	39.80	141	68	48.22
STYLE	6	2	33.33	113	73	64.60

Table 3: The distribution of different edit intentions. # **Edits** indicates the total number of applied edits under the current edit intention, # **Accepts** means the total number of edits accepted by users under the current edit intention, and **% Accepts** is calculated by dividing the total accepted edits with the total applied edits.

Expériences 3/3

Comparaison de la qualité des documents finaux produit avec et sans humain dans la boucle

	Avg. Depths	# Edits	Quality
SYSTEM-HUMAN (ours)	2.5	148	0.68
SYSTEM-ONLY	2.8	175	0.28

Q3: Table 5: Quality comparison results of final revised documents with and without human-in-the-loop. **Avg. Depths** indicates the average number of iterations conducted by the system, **# Edits** means the total number of accepted edits by the system, and **Quality** represents the human judgements of the overall quality of system-revised final documents.

Discussion et travaux futurs

- Créer un modèle qui peut **apprendre des stratégies différentes en fonction de l'itération** où il est rendu
- Améliorer les suggestions de modifications des **catégories qui ont un score faible score**
- **Développer le rôle de l'humain** dans son interaction avec le modèle (“ *such as asking users to re-write the machine-revised text*”)
- Mener une **étude à grande échelle** pour obtenir des **résultats statistiques plus significatifs** (“*e.g. optimal number of revision depths and edit suggestions*”)