Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques

Baptiste Blouin, Benoit Favre, Jeremy Auguste

27-06-2022

ENP-China

“ Le projet "Elites, Networks and Power in modern urban China" explore le processus de transformation des élites en Chine entre 1830 et 1949. Il se concentre sur trois grandes zones urbaines qui ont été les moteurs du changement dans la Chine moderne : Beijing/Tianjin, Guangzhou/Hong Kong, et Shanghai. “

- Plus d'un siècle de presse Chinoise et Anglaise. (∼ 13M documents).
- Nécessité d'utilisation de systèmes d'extraction d'information.
- L'interprétation des résultats nécessite une stabilité au niveau des résultats.


The North-China Herald

AND SUPREME COURT AND CONSULAR GAZETTE

WEEKLY EDITION OF THE NORTH-CHINA DAILY NEWS

Registered at the Chinese P. O. for transmission as mail matter under Class B.

(Tsz-ling-sing-chi-chow-kan 刊週期星林字)

Vol. CLIV. No. 3007. SHANGHAI, SATURDAY, MARCH 28, 1925. Taels 12 per Annum.

CONTENTS

DINNER PARTY

DINNER PARTY

LITTLE HSU'S ADVENTURES IN FRANCE

ANFU

PEKING'S FEAR OF WU PEI-FU

Said to Be Collecting His Old Officers at Yochow: Chinese Prophesying New Move

From Our Own Correspondent.

Peking, Mar. 17.

Wu Pei-fu has sent an urgent message to all his old officers to report at once to Yochow. Several ex-officers have already responded to the call, and it is said that the old Chihli leader has concentrated about two thousand of his old lieutenants at his new headquarters, which is his old stamping-ground. Officials here are apprehensive of another attempt by Wu to get back to power, and the Chinese freely predict his rapid return to the stage of events.

Peking, Mar. 17.

According to messages received from Ichang, General Wu Tao-yuan,

"STRONG MAN" WORK IN PEKING

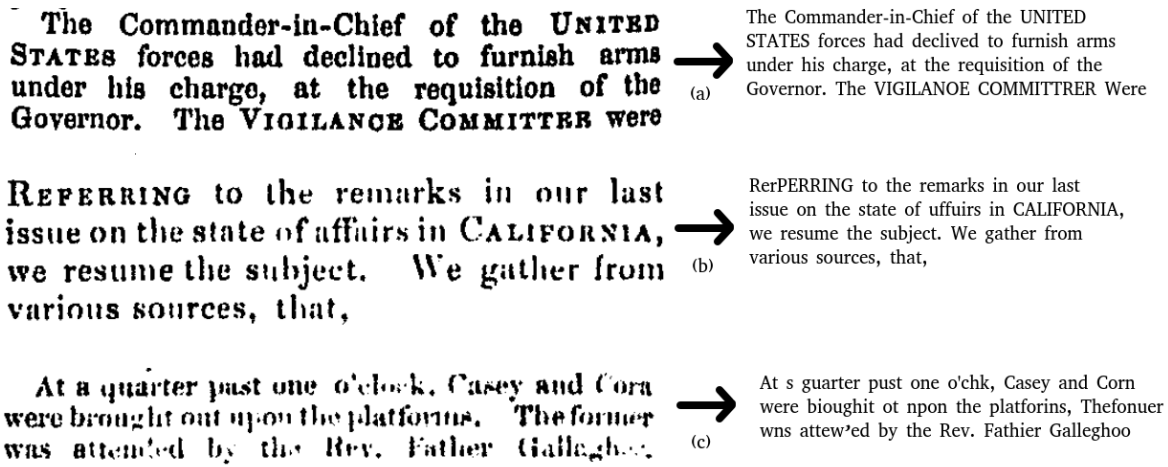
The New Minister of Education Not Deterred from Office by Student Hostility

From Our Own Correspondent.

Peking, Mar. 18.

Tuan Chi-jui has provided at least one example of strong man government in connexion with the assumption of office of Wang Chiu-ling, the Minister of Education, which the incumbent had postponed until after the death of Sun Yat-sen. The acting Minister, Ma Chi-lun, had held the office so long that he was unwilling to retire into the comparative obscurity of the Vice-Ministership. This is no doubt the basis of the allegation that it was he who incited a gang of students to stage a demonstration outside the Ministry when Mr. Wang arrived; they did their work so well that the new Minister could not enter the

Les erreurs d'OCR



- À quel point les erreurs d'OCR impact les tâches de TAL ?
  - Reconnaissance d'entités nommées [Grover et al., 2008, Packer et al., 2010]
  - Étiquetage en partie de discours [Lin, 2003, Mieskes and Schmunk, 2019]
  - Résumé automatique [Jing et al., 2003]
  - [Linhares Pontes et al., 2019, Boros et al., 2022]
- Est-ce que les modèles clef en main sont utilisables sur ce type de données ? [Rodriguez et al., 2012]
- Est-ce un problème de modèle ou un problème des données d'apprentissage ?

Annotations

- Un faible quantité de données annotés pour le domaine historique.
  - Quaero Old Press [Rosset et al., 2012]
  - HIPE [Ehrmann et al., 2020]
  - NewsEye [Hamdi et al., 2021]
- Comment tirer parti des ressources contemporaines annotées ?

Proposition

- Évaluation des nouvelles architectures *Transformers* sur ces difficultés.
- Injection d'erreurs d'OCR dans les données annotées contemporaines.

Configuration

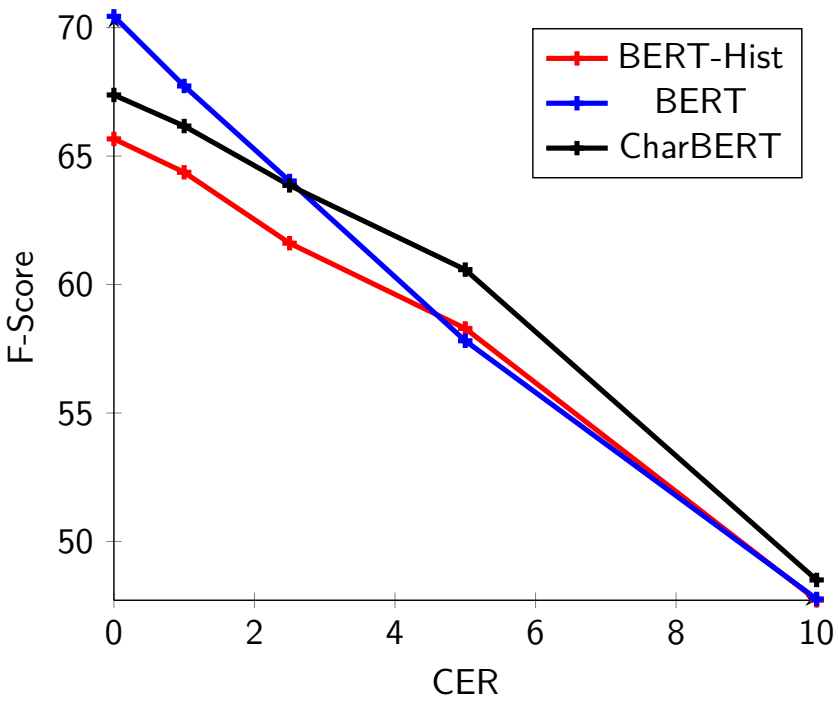
- Dataset
  - **OntoNote 05** [Weischedel et al., 2013]
  - **LitBank** [Bamman et al., 2019]
  - **ACE 05** [Walker et al., 2006]
  - Le Wall Street Journal du jeu de données Penn Treebank [Marcus et al., 1993]
- Tache
  - Reconnaissance d'entité nommée
  - Étiquetage en partie de discours
  - Extraction d'arguments d'événement
- Modèles
  - **BERT** [Devlin et al., 2019]
  - **BERT-Hist** [Hosseini et al., 2021].
    - Microsoft British Library de 1760 à 1900
  - **CharBERT** [Ma et al., 2020]

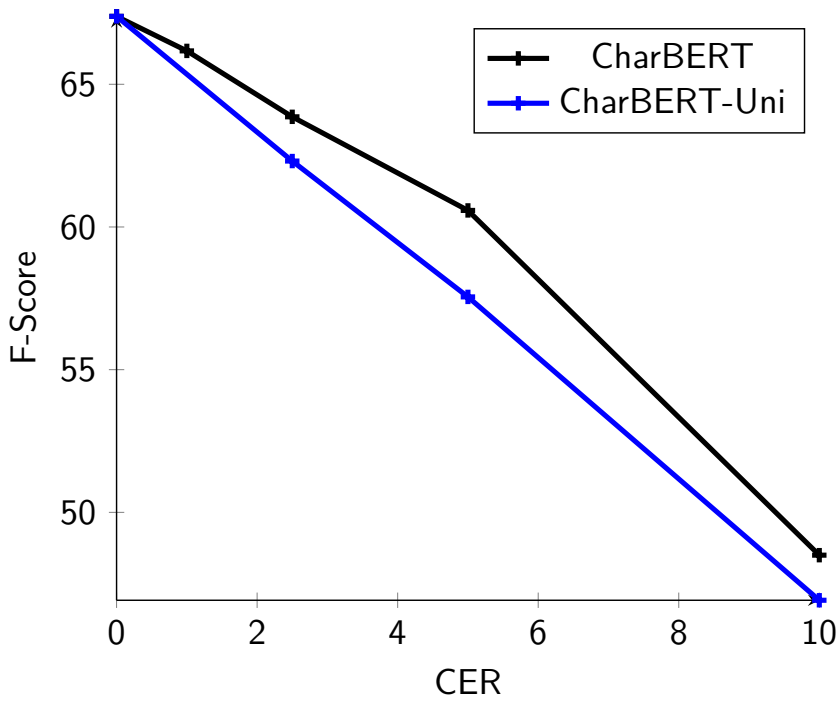
Injection d'erreurs d'OCR

- Modification de la séquence de caractères originale en utilisant la suppression, l'insertion et la substitution [Pruthi et al., 2019].
- Utilisation des probabilités d'erreurs provenant de ICDAR-17 [Nayef et al., 2017].
  - Presses anglaises, de la British Library, allant de 1744 à 1911.
  - Permet de contrôler la quantité de CER à injecter.
- Une injection d'erreurs dans les jeux de tests permet d'évaluer la robustesse des modèles dans un scénario d'adaptation de domaine.

Quantification de l'impact des erreurs d'OCR

Reconnaissance d'entités nommées : ACE → LitBank

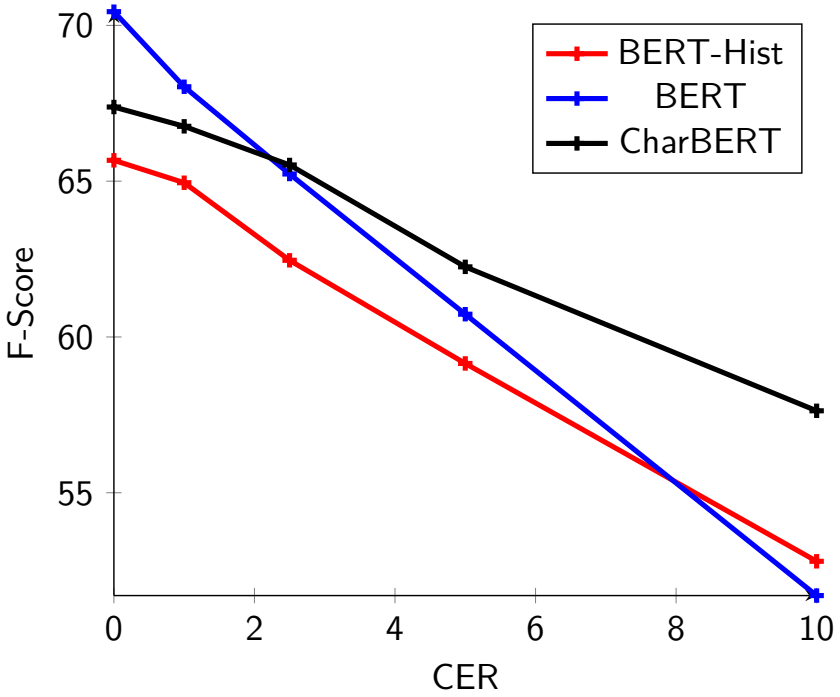


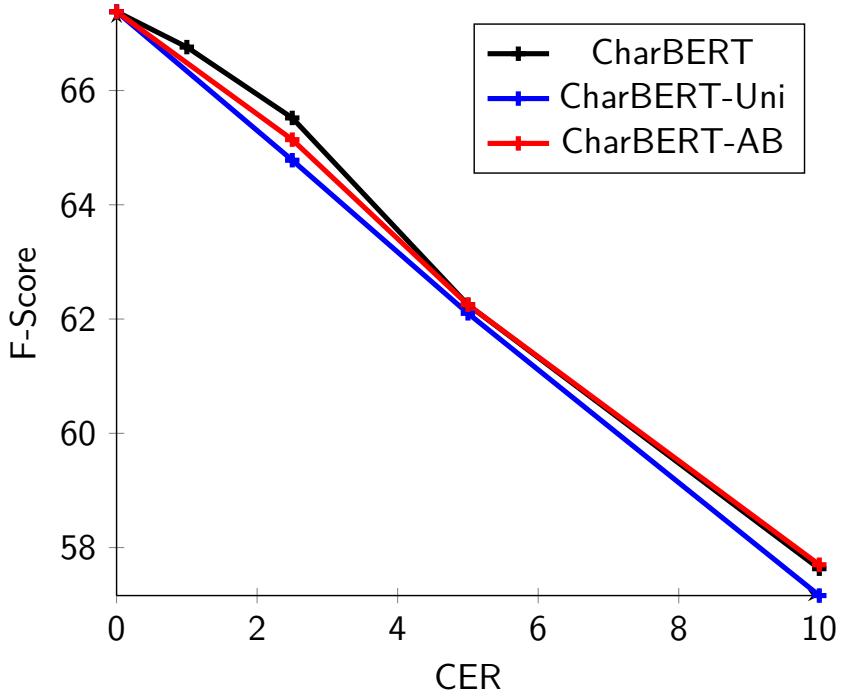


- Atténuation de l'impact du bruit avec les représentations de mots historiques ou avec l'utilisation de CharBERT.

Réduction de l'impact des erreurs d'OCR

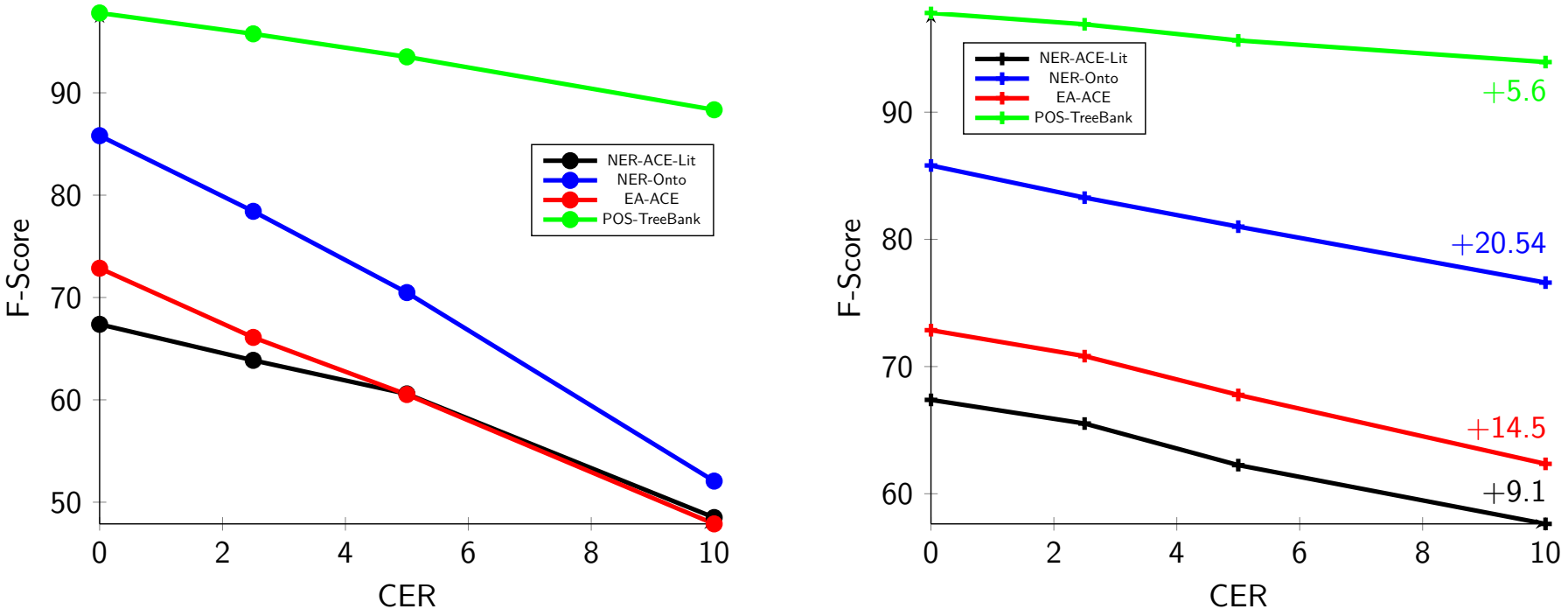
Reconnaissance d'entités nommées : ACE → LitBank





- CharBERT est plus robuste que BERT lors d'un apprentissage sur des données bruitées (+9.1% comparé à +3.94 %).
- La quantité de CER a plus d'impact que la distribution d'erreurs.

Extraction d'informations



- Scénario d'application dans le même domaine.
- Toutes les tâches sont impactées par les erreurs d'OCR.

- Une quantité de bruit non homogène dans les documents historiques.

Extraction d'arguments d'événements : ACE → ACE

Train \ Test	0%	2.5%	5%	10%
0%	**72.86%**	67.12%	63.16%	49.72%
2.5%	68.47%	**70.82%**	66.65%	56.71%
5%	68.30%	68.79%	**67.77%**	59.15%
10%	67.36%	68.58%	67.22%	**62.38%**

- Plus la quantité de bruit est proche de celle des documents d'inférences, plus les résultats sont élevés.
- Un entraînement sur des données bruitées pour le traitement de données non bruitées a moins d'impact que l'inverse.

Conclusion

- Évaluation de l'impact des erreurs d'OCR sur des tâches d'extraction d'informations.
- Amélioration des systèmes de TAL en injectant du bruit dans les données d'entraînement contemporaine.
- La quantité de bruit à plus d'impact que la provenance des erreurs.
- L'utilisation d'embeddings de caractères permet de réduire cet impact.

Travaux futurs

- Correction ciblée de ces erreurs d'OCR sur les informations extraites.
- Analyse d'erreurs approfondie afin de mieux identifier les difficultés liées aux erreurs d'OCR.

Questions ?


Aix*Marseille
université
Socialement engagée


LIS
LABORATOIRE
D'INFORMATIQUE
& SYSTÈMES


ENP
ChINA

References

Bamman, D., Popat, S., and Shen, S. (2019).
An annotated dataset of literary entities.
In *Proceedings of the 2019 Conference of the North*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Boros, E., Nguyen, N. K., Lejeune, G., and Doucet, A. (2022).
Assessing the impact of OCR noise on multilingual event detection over digitised documents.
International Journal on Digital Libraries.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ehrmann, M., Romanello, M., Fluckiger, A., and Clematide, S. (2020).
Extended Overview of CLEF HIPE 2020: Named Entity Processing on