8.8/10    Relecteur 1 (-)    17/04/2024

› Grille de notation

        Pertinence :    4 - Bien

        Clarté :    3 - Moyen

        Correction :    4 - Bien

        Originalité :    3 - Moyen

› Commentaire déposant

        L'article traite de la reconnaissance d'entités nommées (REN) dans le contexte d'oeuvres littéraires dont le texte est issu de reconnaissance optique de caractères (OCR). Il vise à étudier l'effet de la correction (orthographique) automatique des oeuvres ocrisées sur la qualité de la REN effectuée dessus. L'article décrit un ensemble d'expériences sur 3 langues (français, portugais, anglais) avec plusieurs systèmes OCR (kraken, tesseract) et un système de REN off-the-shelf (spacy), combinés ou non avec un correcteur orthographique (Jamspell). Il réalise une étude qualitative et quantitative des différences entre les combinaisons de systèmes (comparaisons des EN trouvées sur texte de ref/texte ocrisé/texte ocrisé avec correction). Le résultat principal de l'étude est que la correction automatique de l'OCR n'améliore pas la REN (car introduit souvent des surcorrections) à l'exception de certains cas (output OCR très bruité)

        Points positifs / points négatifs: cf review du 2eme round.

        Points additionnels:

        L'article s'est beaucoup amélioré sur la description des systèmes utilisés. Par ailleurs, il parvient à mieux synthétiser ses résultats, ce qui donne un article plus lisible. L'article est donc à mon sens publiable dans TAL.

        - sur la typologie des 8 catégories: la présentation est maintenant bien meilleure sans être complètement satisfaisante, il est par ailleurs dommage de présenter cette typologie sans l'utiliser pour quantifier les erreurs dans chaque catégorie

        - il n'est pas clair si le corpus annoté de la section 4.5 va être publié (il me semblait que oui d'après la 2eme version de l'article, mais ce n'est plus mentionnéme semble-t-il), ce serait une contribution intéressante de l'article

        - Tableau 3: les colonnes de pourcentages seraient plus lisibles en les exprimant sous forme    de variation (-5%    au lieu de 95%, +0.7% au lieu de 100.7 %)

        - Il faudrait clarifier ce qui est comparé avec cosinus & jaccard dans la section 5.4.1:    le texte complet de référence vs la version ROC? (expliciter      cela également dans la légende des graphiques). Par ailleurs, je suggère de mettre les comparaisons entre jaccard et cosinus directement sur le git comme les graphiques ne sont pas présentés pour jaccard.

        Réponses à la lettre aux relecteurices:

        - Sur le filtrage des EN-types, mon point était qu'il s'agissait d'une baseline qui aurait eu tout à fait sa place dans l'article

        - Écriture inclusive: la remarque s'est perdue à cause d'un problème d'encodage du point médian (j'imagine à cause de sciences-conf), ma suggestion était simplement d'en utiliser un seul au lieu de 2, c'est-à-dire 'chercheur-euses'/'chercheur?euses' à la place de 'chercheur-se-s'/'chercheur?se?s' (je mets la version avec tiret pour éviter que le problème d'encodage se reproduise, mais en effet je n'ai jamais rencontré d'écriture inclusive avec '?' non plus).

        - après lecture de la feuille de style, il faut juste utiliser \citeasnoun{} ou \cite{} selon le contexte pour résoudre la plupart des problèmes, par ailleurs, utiliser \cite[cf    page 4]{author} pour éviter de juxtaposer des parenthèses. Enfin, \cite{ref1, ref2} pour citer plusieurs articles dans la même paire de parenthèse.

        -> à vérifier    sur l'ensemble de l'article ("harmoniser l'utilisation des citations (auteurice, année) vs auteurice (année) -> les phrases doivent rester grammaticales quand le matériel entre parenthèses est supprimé (cf les instructions dans le template de la revue TAL).")

        - Diagrammes par intersection: si matplotlib ne permet pas d'ajuster les échelles des graphiques, je suggère d'utiliser des rectangles au lieu de cercles pour représenter les intersections, il est beaucoup plus facile de comparer des aires à l'oeil nu sur des rectangles que sur des cercles.

        typos:

        - 4.3: "Kraken-spaCy\_lg). cf. le tableau 9)." parenthèses en trop

        - fin section 4: "Jusque là nous n’avons pas pu trouver de manière convain-

        cante de calculer un F-score et nous commençons à entrevoir les limites de l’usage des

        intersections." -> phrase à reformuler

=====================================

3.8/10    Relecteur 2 (-)    14/05/2024

› Grille de notation

        Pertinence :    4 - Bien

        Clarté :    3 - Moyen

        Correction :    3 - Moyen

        Originalité :    3 - Moyen

› Commentaire déposant

        Les améliorations apportées à l'article ont permis de clarifier divers points et d'en fluidifier la lecture, grâce à des transitions plus claires permettant de mieux articuler les apports de chaque analyse. On comprend désormais mieux l'organisation des diverses expériences et la manière dont elles se complètent. Il reste toutefois encore des points à clarifier (cf. commentaires détaillés)

        La conclusion gagnerait également à être étoffée afin de faire émerger les contributions principales de l'article. En particulier, l'utilisation de JamSpell montre un effet peu probant, mais la conclusion est difficile à généraliser. En effet, elle peut être liée à l'outil particulier utilisé, dont on sait finalement peu de choses et dont le fonctionnement est peu explicité.

        Commentaires détaillés :

        - Le début de la Section 3 reste peu clair, notamment en raison de l'indication "version orthographiquement correcte des textes" qui donne l'impression que les textes de référence ont été corrigés manuellement. Il faufrait explicitement dire que ce n'est pas le cas. Par ailleurs, l'introduction de la section 3 indique "Pour chaque texte nous possédons deux transcriptions différentes", ce qui semblerait indiquer que ces transcriptions automatiques ont été faites en dehors du travail décrit dans l'article. Or, des informations plus détaillées sont données plus tard, dans la section 4.1.1. Il faudrait clarifier le lien entre ce qui est décrit au début de la section 3 et ce qui est présenté dans la section 4.1.1. Si l'on comprend bien, la ROC a été effectuée par vous à partir des PDFs des oeuvres ?

        - L'outil spaCy est utilisé et cité dès la section 3.1 sans avoir été expliqué : il faudrait a minima faire un renvoi vers la section 4.1.2 pour indiquer que l'outil et les configurations utilisées seront détaillés par la suite.

        - La description de spaCy n'est pas toujours claire. Que sont les "lookups data", les "Explosion vectors" ?

        - A quoi sert la ligne spaCy\_lg dans la Table 5, dans la mesure où cette indication se retrouve dans chaque colonne ? Par ailleurs, la Table 4 indique 295 EN identifiées sur Daudet pour spaCy\_lg, mais il n'y en a que 237 dans la table 5. On note la même différence pour les entités nommées dans la référence (209 dans la table 4, 203 dans la table 5). Enfin, les données de la table 5 ne sont pas cohérentes avec la conclusion qui est tirée de ces données : il y a 60 VP hapax identifiées dans la référence mais seulement 57 pour la version ROC Tess. + fr

        - p. 14 vous dites que JamSpell est développé en Python, or ce n'est pas ce qui est indiqué sur https://github.com/bakwc/JamSpell : "multi-language - it's written in C++ and available for many languages with swig bindings"

        - Dans le Tableau 13, est-ce que les lignes MOMC et MOI ELTeC montrent des erreurs dans les textes de références ELTeC ou alors s'agit-il de JamSpell entraîné sur ce corpus ? Ce n'est pas clair.

        - p. 19 : on ne comprend pas la remarque "Après une observation des différentes mesures de distance présentées pour chaque ... cosinus et Jaccard ne mesurent pas la même chose." car les figures 3 et 4 ne présentent que le cosinus et on ne peut donc pas faire de comparaison avec Jaccard. Certes, l'article indique que les résultats sont sur un github, mais ces données ne sont pas disponibles (anonymisation) et on ne peux donc pas s'y référer pour la relecture. On peut donc se poser la question de la pertinence de cette discussion, qui prend beaucoup de place tout en étant très annexe par rapport aux propos de l'article.

        Rédaction et présentation :

        - Résumé : "d'évaluation plus souple\*s\*"

        - Introduction : "de numérisation et de publication des transcriptions (ROC) d’oeuvres" : supprimer (ROC), l'acronyme a déjà été expliqué deux lignes au-dessus

        - Dans l’insertion des références, lorsque la référence bibliographique joue un rôle syntaxique dans le texte, les noms des auteurs ne sont pas entre parenthèses :    "Dupont et Durand (2006) montrent que".

        - Il manque le DOI pour spaCy dans la bibliographie (10.5281/zenodo.1212303)

        - p. 11 : "sont considérées par la machine comme des FP" --> je supprimerais "par la machine"

        - p. 14 : "que sur le globale" -> que sur tous les types d'entités

        - Dans le tableau 9, pourquoi abréger spaCy\_lg par sp ? Il y a largement la place et, d'une manière générale, il vaut mieux privilégier la cohérence dans les noms donnés aux ressources et outils dans un article.

        - p. 17 : "nous avons décider" --> "nous avons décidé"

==========================================

8.8/10    Relecteur 3 (-)    15/05/2024

› Grille de notation

        Pertinence :    4 - Bien

        Clarté :    3 - Moyen

        La clarté a été très fortement améliorée. Il reste juste quelques points qui me semblent moins clairs que je détaille ci-dessous.

        Correction :    4 - Bien

        Originalité :    3 - Moyen

› Commentaire interne

        Le but de l’article est d’analyser l’effet des erreurs de l’OCR sur la reconnaissance d’entités nommées (REN) de lieu dans les corpus littéraires (avant et après correction automatique des erreurs d’OCR).    L’analyse a été effectuée en français, anglais et portugais sur des textes issus de la Très Grande Bibliothèque et des textes des collections française, anglaise et portugaise de la collection européenne ELTeC. Les outils comparés sont deux outils d’OCR: Tesseract et Kraken et un outil de REN: spacy. L’outil de correction automatique utilisé est Jamspell (le modèle pré-entraîné et un autre entraîné sur une partie d’ELTeC).

        Il s’agit d’un troisième lecture de cet article, et globalement l’article a été largement amélioré sur plusieurs points, y compris sur sa structure, qui est beaucoup plus logique, et sur l’utilisation d’un seul outil de REN (en enlevant stanza), ce qui permet de focaliser sur les différents d’outils d’OCR. Plus de détails ont été donnés sur les outils utilisés, et l’état de l’art a été légèrement étendu.

        Il reste encore quelques problèmes à mon avis, ce que je détaille ci-dessous :

                Evaluation souple vs. stricte pour ref.: les résultats diffèrents pour ref. entre l’évaluation stricte et souple, ce qui peut être expliqué par le fait que pour la référence aussi “souple” prend en compte des matches partielles. Cependant, le fait que Tesseract a de meilleurs résultats que la référence selon l’évaluation souple n’indique-t-il pas un problème de sur-estimation des scores avec l’évaluation souple ?

                Jaccard vs. Cosinus: Je trouve cette section étrange. Les résultats de Jaccard ne sont pas donnés, mais ils sont commentés dans le texte (un lien vers un GitHub anonymisé est donné). A mon avis il faut soit les inclure dans l’article, soit enlever la mention des résultats Jaccard. Il est écrit que “il semble que les métriques cosinus et Jaccard ne mesurent pas la même chose” - si j’ai bien compris ce qui a été fait, Jaccard est une mesure de similarité et cosinus de distance, ce qui explique pourquoi les scores cosinus sont autour de 0 et les scores Jaccard autour de 1, donc ce commentaire me semble un peu étrange dans ce contexte. Il serait bien d’expliciter les calculs faits.

        Commentaries plus mineurs:

                Description des outils d’OCR: quand vous parlez de “modèle de langue neuronaux”, est-ce que ça fait référence au modèle d’OCR ou d’une composante du modèle ? Ca mérite d’une description.

                Section 4.2 “tableau 4… des versions de ROC” - il me semble qu’on n’y voit que le système Tesseract pour les trois langues. Si c’est le cas “versions” me semblent un peu inadapté, car ça me fait penser à différents types de systèmes.

                Discussion sur l’augmentation du nombre d’hapax -> est-ce que le nombre d’hapax est donné quelque part ou il est juste induit des autres chiffres?

                Typologie des erreurs: la typologie a été simplifiée, ce que j’apprécie. Cependant, je ne suis toujours pas convaincue de l’utilité de définir les catégories Faux {VP,VN,FP,FN}. Je comprends mieux l’intérêt maintenant d’évoquer ce type d’erreurs, mais je pense que ça pourrait être décrit de façon un peu plus intuitif en évoquant avec quelques détails de plus le fait que chaque cas pourrait être “faux” s’il y a une erreur dans la référence, ou des informations manquantes dans la référence, plutôt que de formaliser chaque type. Par exemple, si j’ai bien compris, Faux vrai positif indique un vrai positif qui n’aurait pas dû l’être, car il s’agit d’une erreur dans la référence, et qui aurait dû être en fait un faux positif.

                Footnote 30: Je ne mettrais pas que vous avez annoté seulement une partie du corpus en raison du temps limité. Il s’agit de l’annotation d’un échantillon du corpus, ce qui est totalement raisonnable et justifiable en soi (sans parler du temps limité, ce qui donne l’impression que c’est pour s’excuser).

        Language, formatting:

                Les citations sont toujours mal formatées, e.g. doubles parenthèses autour de plusieurs citations. La feuille de style donne des exemples sur comment citer plusieurs travaux \cite{Chmielik\_TAL\_52,Virpioja\_TAL\_52}, ainsi que de faire une citation comme sujet de la phrase \citeasnoun{Lavallee\_TAL\_52} ont démontré (…), donc il n’est pas nécessaire normalement d’utiliser d’autres paquets. Il y a aussi un problème avec les italiques, qui sont maintenant soulignés dans le texte (e.g. le “et al” des citations). Il s’agit peut-être d’un problème d’ajout de paquet latex qui créé ce problème.

                Page 10: une entités -> une entité

                Footnote 31: possible espace supplémentaire avant

                4.4: “EN détectées dans les versions de ROC” -> par “versions”, est-ce que vous voulez dire entre la version produite par le système et la référence?

                4.5: , ce qui est, -> enlever la virgule finale

                Le tableau 9 ne suit pas le même formatage que les autres tableaux

                5.1: à mon avis il suffit d’une courte phrase pour décrire pourquoi la configuration portugaise n’est incluse, et le tableau 9 pourrait être enlevé.

                Parfois “ELTeC” est utilisé pour le modèle Jspll entraîné sur ELTeC parfois Jspll-ELTeC - il serait plus clair de les homogénéiser (je préfère le nom plus explicite).

                Figure 3d: .png dans la caption

                Figure 3: les différences sont peu visibles, car les chiffres sont tous près de 0 et l’échelle est très grande. Je suis par ailleurs d’accord avec l’un des autres reviewers, que les échelles devraient être les mêmes pour chaque groupe de figures (peut-être avec la fonction set\_ylim()?). Je vous conseille de consulter d’autres personnes pour demander des conseils si vous n’avez pas encore trouvé comment faire ça dans matplotlib.

› Commentaire déposant

        Le but de l’article est d’analyser l’effet des erreurs de l’OCR sur la reconnaissance d’entités nommées (REN) de lieu dans les corpus littéraires (avant et après correction automatique des erreurs d’OCR).    L’analyse a été effectuée en français, anglais et portugais sur des textes issus de la Très Grande Bibliothèque et des textes des collections française, anglaise et portugaise de la collection européenne ELTeC. Les outils comparés sont deux outils d’OCR: Tesseract et Kraken et un outil de REN: spacy. L’outil de correction automatique utilisé est Jamspell (le modèle pré-entraîné et un autre entraîné sur une partie d’ELTeC).

        Il s’agit d’un troisième lecture de cet article, et globalement l’article a été largement amélioré sur plusieurs points, y compris sur sa structure, qui est beaucoup plus logique, et sur l’utilisation d’un seul outil de REN (en enlevant stanza), ce qui permet de focaliser sur les différents d’outils d’OCR. Plus de détails ont été donnés sur les outils utilisés, et l’état de l’art a été légèrement étendu.

        Il reste encore quelques problèmes à mon avis, ce que je détaille ci-dessous :

                Evaluation souple vs. stricte pour ref.: les résultats diffèrents pour ref. entre l’évaluation stricte et souple, ce qui peut être expliqué par le fait que pour la référence aussi “souple” prend en compte des matches partielles. Cependant, le fait que Tesseract a de meilleurs résultats que la référence selon l’évaluation souple n’indique-t-il pas un problème de sur-estimation des scores avec l’évaluation souple ?

                Jaccard vs. Cosinus: Je trouve cette section étrange. Les résultats de Jaccard ne sont pas donnés, mais ils sont commentés dans le texte (un lien vers un GitHub anonymisé est donné). A mon avis il faut soit les inclure dans l’article, soit enlever la mention des résultats Jaccard. Il est écrit que “il semble que les métriques cosinus et Jaccard ne mesurent pas la même chose” - si j’ai bien compris ce qui a été fait, Jaccard est une mesure de similarité et cosinus de distance, ce qui explique pourquoi les scores cosinus sont autour de 0 et les scores Jaccard autour de 1, donc ce commentaire me semble un peu étrange dans ce contexte. Il serait bien d’expliciter les calculs faits.

        Commentaries plus mineurs:

                Description des outils d’OCR: quand vous parlez de “modèle de langue neuronaux”, est-ce que ça fait référence au modèle d’OCR ou d’une composante du modèle ? Ca mérite d’une description.

                Section 4.2 “tableau 4… des versions de ROC” - il me semble qu’on n’y voit que le système Tesseract pour les trois langues. Si c’est le cas “versions” me semblent un peu inadapté, car ça me fait penser à différents types de systèmes.

                Discussion sur l’augmentation du nombre d’hapax -> est-ce que le nombre d’hapax est donné quelque part ou il est juste induit des autres chiffres?

                Typologie des erreurs: la typologie a été simplifiée, ce que j’apprécie. Cependant, je ne suis toujours pas convaincue de l’utilité de définir les catégories Faux {VP,VN,FP,FN}. Je comprends mieux l’intérêt maintenant d’évoquer ce type d’erreurs, mais je pense que ça pourrait être décrit de façon un peu plus intuitif en évoquant avec quelques détails de plus le fait que chaque cas pourrait être “faux” s’il y a une erreur dans la référence, ou des informations manquantes dans la référence, plutôt que de formaliser chaque type. Par exemple, si j’ai bien compris, Faux vrai positif indique un vrai positif qui n’aurait pas dû l’être, car il s’agit d’une erreur dans la référence, et qui aurait dû être en fait un faux positif.

                Footnote 30: Je ne mettrais pas que vous avez annoté seulement une partie du corpus en raison du temps limité. Il s’agit de l’annotation d’un échantillon du corpus, ce qui est totalement raisonnable et justifiable en soi (sans parler du temps limité, ce qui donne l’impression que c’est pour s’excuser).

        Language, formatting:

                Les citations sont toujours mal formatées, e.g. doubles parenthèses autour de plusieurs citations. La feuille de style donne des exemples sur comment citer plusieurs travaux \cite{Chmielik\_TAL\_52,Virpioja\_TAL\_52}, ainsi que de faire une citation comme sujet de la phrase \citeasnoun{Lavallee\_TAL\_52} ont démontré (…), donc il n’est pas nécessaire normalement d’utiliser d’autres paquets. Il y a aussi un problème avec les italiques, qui sont maintenant soulignés dans le texte (e.g. le “et al” des citations). Il s’agit peut-être d’un problème d’ajout de paquet latex qui créé ce problème.

                Page 10: une entités -> une entité

                Footnote 31: possible espace supplémentaire avant

                4.4: “EN détectées dans les versions de ROC” -> par “versions”, est-ce que vous voulez dire entre la version produite par le système et la référence?

                4.5: , ce qui est, -> enlever la virgule finale

                Le tableau 9 ne suit pas le même formatage que les autres tableaux

                5.1: à mon avis il suffit d’une courte phrase pour décrire pourquoi la configuration portugaise n’est incluse, et le tableau 9 pourrait être enlevé.

                Parfois “ELTeC” est utilisé pour le modèle Jspll entraîné sur ELTeC parfois Jspll-ELTeC - il serait plus clair de les homogénéiser (je préfère le nom plus explicite).

                Figure 3d: .png dans la caption

                Figure 3: les différences sont peu visibles, car les chiffres sont tous près de 0 et l’échelle est très grande. Je suis par ailleurs d’accord avec l’un des autres reviewers, que les échelles devraient être les mêmes pour chaque groupe de figures (peut-être avec la fonction set\_ylim()?). Je vous conseille de consulter d’autres personnes pour demander des conseils si vous n’avez pas encore trouvé comment faire ça dans matplotlib.