Mapping spatial named entities from noisy OCR output: Epimetheus from OCR to map.

Koudoro-Parfait, Caroline

caroline.parfait@sorbonne-universite.fr
ObTIC, Observatoire des textes des idées et des corpus, Sorbonne Université, France; STIH, Sens Texte Informatique Histoire, Sorbonne Université, France; SCAI, Sorbonne Center for Artificial Intelligence, Sorbonne Université, France

Alrahabi, Motasem

motasem.alrahabi@sorbonne-universite.fr
ObTIC, Observatoire des textes des idées et des corpus, Sorbonne Université, France

Dupont, Yoann

yoa.dupont@gmail.com
Lattice, Langues, Textes, Traitements informatiques, Cognition, France

Lejeune, Gaël

gael.lejeune@sorbonne-universite.fr
STIH, Sens Texte Informatique Histoire, Sorbonne Université, France

Roe, Glenn

glenn.roe@paris-sorbonne.fr
ObTIC, Observatoire des textes des idées et des corpus, Sorbonne Université, France

This contribution presents both the difficulties encountered, and the methods used to overcome them, when using out-of-the-box tools for the elaboration of text-processing pipelines that go from optical character recognition (OCR) to named entity recognition (NER) and the cartographic representation of places mentioned in literary texts. NER's significance in NLP stems from the fact that named entities (NEs) constitute the majority of search queries formulated by users, especially spatial NEs (van Strien, Daniel et al. 2020). Better identification of NEs is thus an efficient way to improve access to data in large textual corpora ((Chiron, Guillaume et al. 2017); (Linhares Pontes, Elvys et al. 2019)).

Researchers are often confronted with the difficulties of applying NLP tools – which are normally trained on clean data (Eshel, Yotam et al. 2017) – to less standardized texts such as OCR transcripts. OCR errors are commonly known as noise : insertions, deletions, and also substitutions of one or more characters by others. To overcome these difficulties, one approach is to automatically correct the OCR output. While for systematic errors automation is possible [1](Stanislawek, Tomasz et al. 2019), for others it becomes difficult to achieve. Automatic correction may even produce new errors (Huynh, Vinh-Namet al. 2020).

To tackle this problem, (Koudoro-Parfait, Caroline et al. 2021) and (Koudoro-Parfait, Caroline et al. 2022) produced a manual and automatic evaluation of the impact of OCR errors on NER. These analyses are based on the French corpus of the European collection of literary texts – ELTeC (Schöch, Christof et al. 2021). The authors compare the NEs obtained on this reference version to the outputs of different noisy OCR transcriptions. They demonstrate that some off-the-shelf NER tools spaCy (Honnibal, Matthew / Montani, Ines 2017)and stanza (Qi, Peng et al. 2020) can be rather robust when it comes to the variations that systems face : for example, contexts and NE contaminated (Hamdi, Ahmed et al. 2022) by different OCR errors. In the same context, (Hamdi, Ahmed et al. 2020) and (van Strien, Daniel et al. 2020) evaluate the impact of OCR noise on subsequent uses of these data. They report that the impact of OCR on NER is rather moderate unless the OCR is really of very poor quality.

As part of the development of the Pandore toolbox (Mayra Cordova, Johanna et al. 2022) by the project team Observatoire des Textes des Idées et des Corpus - ObTIC, we constructed Épiméthée [2]a processing pipeline that allows users to OCR the images of their texts with the French model of Tesseract (Smith, Ray 2007) and to obtain two outputs. The first is a representation of spatial NE's geolocalized on a map. The second is a downloadable file with all the data and metadata that produced the map. Two NER tools spaCy and Flair (Akbik, Alan et al. 2019)are used to produce this output. For each NE, the tool that extracted the NE is indicated, as well as the geolocations automatically generated with the Geopy library (Esmukov, Kostya / Tigas, Mike 2018). The user can modify and correct the generated output and reuse it in the geographic information system of itschoice. The first uses showed that a majority of the NEs that had been recovered by the two NER tools were true positives, even when their form is contaminated. In order to help users to correct NER outputs, we implemented two solutions for the automatic alignment of different variants of a NE. This would allow us to group the different contaminated forms of an NE (for instance, "Saint-Nizier", "Saint-Nizier.n", "Saint-Nizierl", or, "Rhône", "Rhone" et "Rh6ne"). The first is inspired by the work about various distances (Jaccard, Bray-Curtis, Cosine) of (Koudoro-Parfait Caroline et al. 2022), and the second uses a data clustering method ((Lin, Dekang 1998);(Green, Spence et al. 2012)) to get groups of similar NE.

We consider that the clustering can be used as a decision support step (Olteanu, Alexandru L. 2013), in a use case where a user would seek to sort NEs more efficiently than occurrence-by-occurrence or even shape-by-shape processing. Ideally, the user would proceed as follows : (i) a cluster whose centroid is not an NE can be left out, (ii) a cluster whose centroid is indeed an NE can be pre-selected before being re-filtered by the user. The clustering of textual data provides a quantitative and condensed data representation (Loustau, Sébastien 2013) and could be applied to noisy data (Brunet, Camille / Loustau, Sébastien 2013). For clustering, we used the Affinity Propagation algorithm (Frey, Brenda J. / Dueck, Delbert 2007), which is interesting because it does not require a predetermined number of expected clusters. We configured the tool and the cluster method seems efficient, even if it still raises some difficulties, in a few cases the centroid is not a NE but the cluster contains NEs. Our contribution will be thus twofold : we will first present the Épiméthée tool that will be made available to the DH community and second, present an analysis of the added value for users demonstrated in our different use-cases on a corpus of 19th century French novels (Koudoro-Parfait, Caroline et al. 2021).

Notes

1. Substitution of an "é" with an "e"
2. Prototype of Épiméthée : http://pp-obtic.sorbonne-universite.fr/toolbox/ocr_map [30.04.2023].

Bibliography

Akbik, Alan / **Bergmann, Tanja** / **Blythe, Duncan** / **Rasul, Kashif** / **Schweter, Stefan** / **Vollgraf, Roland** (2019): *"FLAIR: An easy-to-use framework for state-of-the-art NLP"*,in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59.

Brunet, Camille / **Loustau, Sébastien** (2013): *"The algorithm of noisy k-means"*, in: CoRR, abs/1308.3314.

Burnard, Lou / **Schöch, Christof** / **Odebrecht, Carolin** (2021): *"In Search of Comity: TEI for Distant Reading"*, in: Journal of the Text Encoding Initiative 14. DOI: https://doi.org/10.4000/jtei.3500 [04.05.2023].

Chiron, Guillaume / **Doucet, Antoine** / **Coustaty, Mickaël** / **Visani, Muriel** / **Moreux, Jean-Philippe** (2017): *" Impact of OCR errors on the use of digital libraries Towards a better access to information "*, in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, Canada. IEEE.

Eshel, Yotam / **Cohen, Noam** / **Radinsky, Kira** / **Markovitch, Shaul** / **Yamada, Ikuya** / **Levy, Omer** (2017): *"Named entity disambiguation for noisy text"*, in: Levy, Roger /Specia, Lucia, editors, the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017, pages 58–68. Association for Computational Linguistic.

Esmukov, Kostya / **Tigas, Mike** (2018): https://geopy.readthedocs.io/en/stable/[30.04.2023].

Frey, Brenda J. / **Dueck, Delbert** (2007): *"Clustering by passing messages between data points"*, in: Science, 315(5814):972–976.

Green, Spence / **Andrews, Nicholas** / **Gormley, Matthew R.** / **Dredze, Mark** / **Manning, Christopher D.** (2012): *"Entity clustering across languages"*, in: the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 60–69, Montréal, Canada. Association for Computational Linguistics.

Hamdi, Ahmed / **Jean-Caurant, Axel** / **Sidère, Nicolas** / **Coustaty, Mickaël** / **Doucet, Antoine** (2020): *"Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition"*in: Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, pages 87–101.

Hamdi, Ahmed / **Linhares Pontes, Elvys** / **Sidère, Nicolas** / **Coustaty, Mickaël** / **Doucet, Antoine** (2022): *"In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking"*in: Natural Language Engineering. 29(2):425–448. DOI: 10.1017/S1351324922000110.

Honnibal, Matthew / **Montani, Ines** (2017): *"spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing"*, 7(1):411–420.

Huynh, Vinh-Nam / **Hamdi, Ahmed** / **Doucet, Antoine** (2020): *"When to Use OCR Post-correction for Named Entity Recognition?"*in: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, pages 33–42.

Koudoro-Parfait, Caroline / **Lejeune, Gaël** / **Roe, Glenn** (2021): *"Spatial named entity recognition in literary texts: What is the influence of OCR noise?"*, in: Moncla, Ludovic / Brando, Carmen /McDonough, Katherine, editors, GeoHumanities@SIGSPATIAL 2021: the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021, pages 13–21. ACM. https://github.com/These-SCAI2023/NER_GEO_COMPAR [30.04.2023],

Koudoro-Parfait, Caroline / **Lejeune, Gaël** / **Buth, Ritchy** (2022): *"Resolution of entity linking issues on noisy ocr output: automatic disambiguation tracks"*, in: Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN), pages 45–55.

Lin, Dekang (1998): *"Automatic retrieval and clustering of similar words"*, in: COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics.

Linhares Pontes, Elvys / **Hamdi, Ahmed** / **Sidère, Nicolas** / **Doucet, Antoine** (2019): *"Impact of OCR Quality on Named Entity Linking"*, in: International Conference on Asia-Pacific Digital Libraries 2019, Kuala Lumpur, Malaysia.

Loustau, Sébastien (2013): *"Anisotropic oracle inequalities in noisy quantization"*, https://arxiv.org/abs/1305.0630[27.04.2023].

Mayra Cordova, Johanna / **Dupont, Yoann** / **Petkovic, Ljudmila** / **Gawley, James** / **Alrahabi, Motasem** / **Roe, Glenn** (2022): *"Toolbox : une chaîne de traitement de corpus pour les humanités numériques"*, in: Estève, Y., Jiménez, T., Parcollet, T., and Zanon Boito, M., editors, Traitement Automatique des Langues Naturelles, pages 11–13, Avignon, France. ATALA. http://pp-obtic.sorbonne-universite.fr/toolbox/ [30.04.2023].

Olteanu, Alexandru L. (2013): *"On clustering in multiple criteria decision aid: theory and applications"*, in: PhD, Télécom Bretagne, Université de Bretagne Occidentale.

Qi, Peng / **Zhang, Yuhao** / **Zhang, Yuhui** / **Bolton, Jason** / **Manning, Christopher D.** (2020): "Stanza: A python natural language processing toolkit for many human languages", in: Celikyilmaz, A. and Wen, T., editors, the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, pages 101–108. Association for Computational Linguistics.

Schöch, Christof / **Patraș, Roxana** / **Santos, Diana** / **Erjavec, Tomaž** (2021): "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives", in: Modern Languages Open 1/25. DOI: http://doi.org/10.3828/mlo.v0i0.364 [04.05.2023].

Smith, Ray (2007): *"An overview of the tesseract ocr engine"*, in: Ninth international conference on document analysis and recognition (ICDAR 2007), volume 2, pages 629–633. IEEE. https://github.com/tesseract-ocr/tesseract [04.05.2023].

Stanislawek, Tomasz / **Wróblewska, Anna** / **Wójcicka, Alicja** / **Ziembicki, Daniel** / **Biecek, Przemyslaw** (2019): "*Named entity recognition - is there a glass ceiling?"*, in: the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 624–633.

Strien, Daniel van / **Beelen, Kaspar** / **Ardanuy, Mariona** / **Hosseini, Kasra** / **McGillivray, Barbara** / **Colavizza, Giovanni** (2020): *"Assessing the Impact of OCR Quality on Downstream NLP Tasks"*, in, the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH, pages 484 – 496.