NEWS EYE

A Digital Investigator for Historical Newspapers

Dans le cadre de ce projet, des chercheurs en sciences humaines, des informaticiens et des bibliothécaires travaillent ensemble sur la presse historique conservée dans trois bibliothèques nationales (Autriche, Finlande, France).

https://www.univie.ac.at/newseye/

www.newseye.eu

@newseyeeu

NEWS EYE

Enrichissement sémantique de contenus océrisés

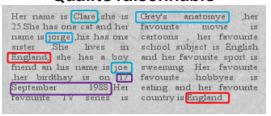
Contact :
Antoine Doucet
antoine.doucet@univ-lr.fr

L'enrichissement sémantique de corpus textuels consiste en l'analyse de documents et l'ajout de données sémantiques à leurs contenus. Parmi une palette de données sémantiques, nous nous concentrons sur les entités nommées et, plus précisément, nous visons à reconnaître ces entités dans les documents, à les désambiguïser via une base de connaissances. De plus, les extraits de texte dans lesquels elles figurent leur sont associés.

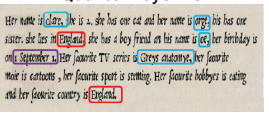
Cependant, le texte des documents océrisés contient de nombreuses erreurs, du fait du vieillissement des documents, de leurs conditions de stockage et/ou de la qualité médiocre du support d'impression originel. Ces erreurs réduisent les performances de tous les traitements ultérieurs en langage naturel (tels que la reconnaissance des noms de personnes, de lieux, d'organisations, et l'analyse d'opinion et de sentiment).

Afin de surmonter ces difficultés, NewsEye conçoit des approches basées sur les techniques les plus récentes (réseaux de neurones artificiels et apprentissage profond) résistantes aux erreurs de la reconnaissance automatique du texte et indépendantes de tout langage. Sur 13 concurrents, les méthodes développées pendant le projet NewsEye ont atteint la première place dans 50 des 52 classements pour la reconnaissance des noms de personnes, de lieux et d'organisations en anglais, français et allemand au concours CLEF HIPE 2020.

Qualité raisonnable



Qualité moyenne



Mauvaise qualité