NEWS EYE

A Digital Investigator for Historical Newspapers

In this project, humanities researchers, computer scientists, and librarians work together on historical newspapers of three national libraries (Austria, Finland, France)

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

www.newseye.eu

@newseyeeu

NEWS EYE

Semantic Text Enrichment

for OCRed content

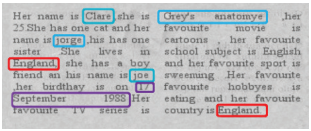
Contact:
Antoine Doucet
antoine.doucet@univ-lr.fr

Semantic text enrichment consists of analyzing documents and adding semantic metadata to their contents. Among several semantic metadata, we concentrate on named entities and, more precisely, we aim to recognize these entities in documents and disambiguate them to a knowledge base. In addition, these entities will be associated with the stance of the text in which they are mentioned.

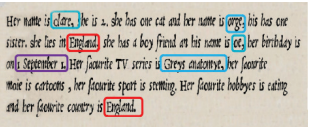
However, OCRed documents contain numerous errors due to the state of documents, as a result of aging, poor storage conditions and/or the low quality of initial printing materials. These errors reduce the performance of all downstream natural language processing tasks (such as NER, NEL and stance detection).

To overcome these problems, NewsEye designs approaches based on the most recent techniques (neural networks and deep learning) that are robust to OCR problems and language independent. Out of 13 competitors, NewsEye approaches reached first place in 50 of 52 leaderboards on NER and NEL in English, French and German at the CLEF HIPE 2020 competition.

Reasonable quality



Medium quality



Poor quality