Extended Named Entity Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign

Olivier Galibert[†] **Sophie Rosset**[*] **Cyril Grouin**[*]
Pierre Zweigenbaum[*] **Ludovic Quintard**[†]

[*]LIMSI-CNRS
91403 Orsay – France
{firstname.lastname}@limsi.fr

[†]LNE
78197 Trappes – France
{firstname.lastname}@lne.fr

Abstract

Within the framework of the Quaero project, we proposed a new definition of named entities, based upon an extension of the coverage of named entities as well as the structure of those named entities. In this new definition, the extended named entities we proposed are both hierarchical and compositional. In this paper, we focused on the annotation of a corpus composed of press archives, OCRed from French newspapers of December 1890. We present the methodology we used to produce the corpus and the characteristics of the corpus in terms of named entities annotation. This annotated corpus has been used in an evaluation campaign. We present this evaluation, the metrics we used and the results obtained by the participants.

Keywords: Named Entity, Press Archives Annotation, Evaluation.

1. Introduction

The evaluation of named entity recognition (NER) methods is an active field of research. NER can be performed on textual data such as newspapers, spoken data or digitized data. Here we focus on French 'old press' data composed of digitized 19th century newspapers.

There has been various work on named entity detection in historical data (Miller et al., 2000; Crane and Jones, 2006; Byrne, 2007; Claire Grover and Ball, 2008).

In terms of evaluation campaigns, ACE (Doddington et al., 2004) included NER on OCRed data; for French, the Quaero program organized a first evaluation campaign of NER on 'old press' data (Galibert et al., 2010).

In this paper we present an annotation and evaluation campaign on Extended Named Entity recognition in French OCRed 'old press' which was organized within the Quaero program. It followed an evaluation campaign on spoken data (Galibert et al., 2011) but presents many specific characteristics. Quaero[1] is a program promoting research on and industrial innovation in technologies for automatic analysis and classification of multimedia and multilingual documents.

Section 2. presents the definition of the Quaero extended named entities. Section 3. describes the OCR data used for training and test in this campaign and the pre-processing steps applied to these data. The evaluation campaign along with metrics used and results obtained by the different participants is then presented in Section 4. Finally, we conclude with some discussions and perspectives in Section 5.

2. Quaero Named Entity definition

In this section, we briefly define the Quaero extended named entities and contrast them with previous work, present their scope and their hierarchical and compositional nature.

[1] http://www.quaero.org

2.1. Named Entity Types

Initially, Named Entity recognition was described as recognizing proper names (Coates-Stephens, 1992). Since MUC-6 (Grishman and Sundheim, 1996), named entities encompass three major classes: *person, location* and *organization*. Some numerical types are also often described and used in the literature: *date, time* and *amount* (money and percentages in most cases).

Proposals have been made to sub-divide existing categories into finer-grained classes, e.g. *politician* as part of the *person* class (Fleischman and Hovy, 2002) and *city* in the *location* class (Fleischman, 2001). New classes have been added during the CONLL conference. More recently, larger extensions have been proposed: *product* by (Bick, 2004) while (Sekine, 2004) defined an extensive hierarchy of named entities containing about 200 types. In more detailed projects, such as the Virginia Banks project (Crane and Jones, 2006), specific categories were added to fit the considered period (the American Civil War): *ships*, *regiments*, and *railroads* for example.

2.2. Scope

As we aimed to build a fact database from news data, we focused on the extraction of entities and relations. We extended the coverage of named entities not only by subdividing the existing classes, as has been done in the aforementioned work, but also by supporting new kinds of entities. The extended named entities we defined are both hierarchical and compositional. This structure requires novel methods to evaluate system outputs. Compared to existing named entity definitions, our approach is more general than the extensions proposed for specific domains, and is simpler than the extensive hierarchy defined by Sekine (2004). This structure allows us to cover a large number of named entities with a basic categorization which provides a foundation which facilitates further annotation work. The guidelines are available online (Rosset et al., 2011).

2.3. Hierarchy

We used two kinds of elements to define a named entity: types and components. *Types and subtypes* refer to a general segmentation of the world into major categories. We consider that structuring the contents of an entity is important too: within these categories, we defined a second level of annotation we call *components*. A comprehensive description of both types and components can be found in (Grouin et al., 2011).

Types and subtypes refer to the general category of a named entity. They constitute the first level of annotation and give general information about the annotated expression. The taxonomy is composed of 7 types (*person, location, organization, amount, time, production* and *function*) and 32 sub-types (individual person *pers.ind* vs. group of persons *pers.coll*, absolute date *time.date.abs* vs. relative date *time.date.rel*, etc.).

Components can be considered as clues to make annotations: either to determine the named entity type (a first name is a clue for the individual person *pers.ind* subtype), or to set the named entity boundaries (a given token is a clue for the named entity, and is within its scope, while the next token is not a clue and is outside its scope). Components are second-level elements, and can never be used outside the scope of a type or subtype element.
An entity is thus composed of components. Two kinds of components are found:

- Transverse components can fit within any type of entity (*name, kind, qualifier, demonym, val, unit, object* and *range-mark*).
- Specific components are only used in a reduced set of components (for example, *name.last, name.first, name.middle* and *title* are only used as components of the *pers.ind* sub-type).

2.4. Structure

In line with the Ester II evaluation campaign (Galliano et al., 2009), we catered for a structured (or 'nested') nature of extended named entities.[2] We defined three kinds of structuring:

1. a type contains a component: the *pers.ind* type (individual person) contains several components such as *title* and *name.last* (Figure 1);
2. a type includes another type, used as a component. In Figure 1, the *func.ind* type (individual function), which spans the whole expression *minister of war*, includes the *org.adm* type (administrative organization), which spans the single word *war*;
3. in cases of metonymy and antonomasia, a type of entity is used to refer to another type of entity (Figure 2). The type to which the entity intrinsically belongs is annotated (the *loc.adm.nat* type, a national administrative location). This entity type is over-annotated with the type to which the expression belongs in the considered context (the *org.adm* type, an administrative organization).

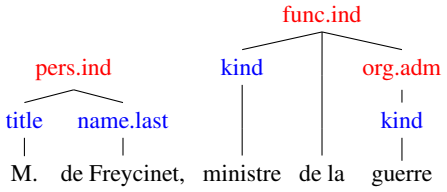

Figure 1: Multi-level annotation of entity types (red tags) and components (blue tags): *Mr de Freycinet, minister of war*

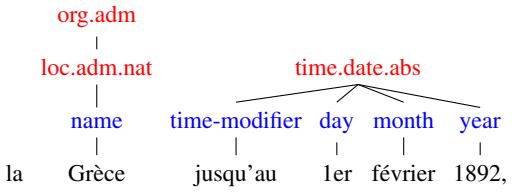

Figure 2: Annotation with types (red tags) and components (blue tags) including metonymy: *Greece, until February 1st, 1892,*

3. Data processing and description

The Quaero Old Press corpus consists in 76 newspaper issues published in December 1890, provided by the French National Library.[3]
We used three different French titles: *Le Temps*, 54 documents for a total amount of 209 pages, *La Croix*, 21 documents for a total amount of 84 pages, and *Le Figaro*, 1 document composed of 2 pages.

3.1. Block selection

A newspaper is composed of various parts (titles, articles, ads, etc.), some of which are not useful for named entity annotation. A corpus study allowed us to determine parts in which we considered annotation would be useless: *titles, mastheads, ads, tables of numbers, theater programmes, stock exchange results, weather reports, etc.*
OCR processing cuts each document page into blocks in which it has recognized text. To filter out irrelevant parts in each page, we designed a block selection tool. This tool, which was developed by LNE,[4] takes as input the image of a page and the standard XML ALTO[5] OCR output for this page, which provides boundary information for each text block. It provides an interface (Figure 3) through which we could interactively select / deselect OCRed blocks of texts.

[2] While compositionality was defined in Ester II, its evaluation did not consider the inclusion covered in the present subsection.

[3] BNF: Bibliothèque Nationale de France.

[4] LNE: Laboratoire National de métrologie et d'Essais.

[5] ALTO: Analyzed Layout and Text Object.

It outputs an XML file where only the selected blocks are kept.

Given a whole newspaper page, this tool links the picture of the page with its corresponding OCRed text in XML format. As outlined above, both picture and text are segmented by the OCR process into segments of various sizes (a word, a sentence, a paragraph). This tool generates separate images for each selected block and one file containing recognized text segments corresponding to each block. Initially, all text segments are selected: the user simply needs to click on irrelevant segments to deselect them. Finally, a file composed of those text segments that were kept selected (blue blocks on figure 3) is produced by this tool. Named entity annotations will be performed on this output text file.

Figure 3: Block selection in a page (selected blocks are in blue, the user has deselected titles and headers).

The first three blocks selected in the picture file (Figure 3) corresponds to three text segments shown in Source 1. The first picture block corresponds to the text from line 2 to line 12 in Source 1 (i.e., the begining of a paragraph), the second picture block corresponds to the sole line 15 (the hyphenation of the last word from the previous paragraph), and the third block corresponds to the text from line 18 to line 22 (a whole paragraph).

3.2. Annotation adaptation

The Quaero Old Press corpus presents the following characteristics. First, the corpus has been created from OCRed press archives; while its estimated OCR quality rate is good,[6] some incorrectly recognized characters remain. Second, the corpus refers to a past period (December 1890):

[6]Each file includes metadata which include its raw OCR quality rate, which ranges from 14.82% to 93.92%, with a median of 82%. Manual corrections were performed after OCR but the quality rate was not updated and is therefore not available to us.

Source 1: First Text Blocks Selection

```
1  00212633/PAG_1_TB000012.png
2  Aujourd'hui, nous donnons comme
3  premier article la reponse magistrale
4  que le cardinal de Paris a faite à une
5  consultation de catholiques, qui n'a pu
6  paraître en notre édition d' hier soir.
7  Ce document trace avec une grande
8  autorité le programme qui fait le fond
9  des revendications de la Ligue de
10 l' "Ave Maria" et dont il faut pour-
11 suivre, ajoute l' Eminent prélat, la réa-
12 lisation "avec calme, énergie et persé-
13
14 00212633/PAG_1_TB000013.png
15 vérance".
16
17 00212633/PAG_1_TB000014.png
18 i Le cardinal termine en disant qu'il
19 faut faire des oeuvres et prier.
20 Nous recevons ce document avec
21 d'autant plus de joie qu'il devient une
22 sorte de consécration de nos efforts.
```

knowledge about named entities from this period is more difficult to obtain accurately, especially for person names. Last, because the original document is a paper edition of a newspaper, the text is formatted into fixed-sized columns; in consequence, the resulting text contains line breaks, with a few hyphenations. In order to take into account these features and to fulfill annotators' requirements, we introduced a new attribute and a new component into the annotation schema:

Attribute *correction*. Annotators have been asked to correct incorrectly recognized entities. To save time and effort, correction must be performed only on named entities, not on non-entity text. In the annotation scheme, the erroneous OCR output for an entity is left in the document, while the corrected entity is inserted in a *correction* attribute.

```
<pers.ind correction="Le Moine">
  <name.last> LE Moibte. </name.last>
</pers.ind>
```

In this example, the last name "Le Moine" has been OCRed as "LE Moibte.". Because this text segment is an entity, the human annotator filled out the correction attribute with the correct name. The correction attribute must be inserted into the regular tag for the (erroneous) entity type.

Component *noisy-entities*. When a character recognition error involves an entity boundary, a segmentation error occurs, either between an entity and other tokens, or between several entities and possibly other tokens. In order to allow the annotator to annotate the entity present in that character span, we defined a new component *noisy-entities* which indicates that an entity is present within the noisy span of characters.

```
<loc.adm.reg correction="EN ALSACE-LORRAINE">
  <noisy-entities>
    KN_ALSACE'LOBR4INE
  </noisy-entities>
</loc.adm.reg>
```

In this example, the text segment "KN_ALSACE'LOBR4INE" includes both recogni-

tion and segmentation errors. The human annotator corrected the segment into "EN ALSACE-LORRAINE" in the *correction* attribute and used the *noisy-entities* component to indicate that within this erroneous segment, an entity "ALSACE-LORRAINE" must be annotated with sub-type *loc.adm.reg* (a regional administrative location, i.e., smaller than a state but larger than a town).

General principles. We defined the following principles to help the human annotators doing annotations on the old press corpus.

- Text portions without any space including at least one entity must be wholly annotated: "rélyséeavaitcoûté22788740fr", "lAlsace-Lorraineque", "Au31janvier";[7]
- Because the original text must not be corrected, only the type or sub-type tag must surround the badly recognized entity; no component tag must be used for this entity (*name, kind,* etc.);
- If a badly recognized segment includes several entities, the *noisy-entities* component must be used: the segment "rélyséeavaitcoûté22788740fr" is composed of entities of two types, a *loc.fac* sub-type (a facility location) for "élysée" and an *amount* sub-type for "22788740fr"; the expected annotation must be:

  <noisy-entities correction="l'élysée avait coûté 22788740 fr">
  rélyséeavaitcoûté22788740fr
  </noisy-entities>

  In this case, no type nor sub-type nor component tag must be used;
- Typographic case errors must be corrected inside the correction attribute;
- Punctuations glued to entities must be left intact; the annotation is performed over the whole entity:

  il est allé à <loc.adm.town> Paris. </loc.adm.town>
  <loc.adm.sup> l'Afrique </loc.adm.sup>

- Nevertheless, punctuation OCR errors must be corrected inside the correction attribute:

  <noisy-entities>
  <pers.ind corr="vicomte de Constantin">
  <title> vicomte </title>
  <name.last> de. Constantin </name.last>
  </pers.ind>
  </noisy-entities>

- In cases of hyphenation and line breaks within a named entity, the annotation must encompass the whole segmented named entity:

  <func.coll> <kind> clientèle </kind> <demonym> ir-landaise, </demonym> </func.coll>

  In this example, the adjective "irlandaise" (*irish*) has been segmented into "ir-" and "landaise" due to hyphenation. The annotation spans the two parts of the entity across two lines.
- Annotations are kept within each block; we did not allow annotations spanning several blocks.

We give in Source 2 the extended named entities annotations performed by the human annotators on the first three blocks of Figure 3.
Due to line breaks, a few annotations encompass two lines, such as the "Ligue de l' "Ave Maria"" extended named entity located from line 9 to line 10.

4. Evaluation campaign

4.1. Organization

Extended Named Entity extraction in old press data was evaluated in the Quaero program in 2011. The process followed a standard profile, with training / development data provided in advance for system development and the evaluation happening afterwards. Training and test corpora are described in Table 1.

Information \ Data	Training	Test
# pages	231	64
# lines	192,543	61,088
# words	1,297,742	363,455
# distinct words	152,655	64,749
# entity types	113,591	35,029
# entity types w. correction	4,122	2,248
# distinct entity types	32	35
# components	166,151	38,495
# components w. correction	204	38
# distinct components	27	24

Table 1: Quaero Old Press training corpus annotated with extended named entities.

The format used in this campaign is a simple text format containing entities marked with XML tags. The input of the systems correspond to the example given in Source 1, and the output (and the training data) follows the form given in Source 2.

4.2. Metrics

Two categories of metrics are usual in the named entity evaluation domain. The first category is the *precision / recall / F-measure* triplet. The principle is simple: count correct annotations and divide them by the number of either reference annotations (recall) or hypothesis annotations (precision); then use harmonic mean to combine both values. This evaluation method is useful, in particular because of its simplicity and associated ease of understanding, but has its limits, especially in noisy input conditions,

[7] Respectively: "theélyséehadcost22788740fr", "theAlsace-Lorrainethat", "OnJanuary31st".

Source 2: Annotated First Text Blocks Selection

```
00212633/PAG_1_TB000012.png
<time.date.rel> <name> Aujourd'hui, </name> </time.date.rel> nous donnons comme
premier article la réponse magistrale
que le <func.ind> <kind> cardinal </kind> de <loc.adm.town> Paris </loc.adm.town> </func.ind> a faite à une
consultation de <pers.coll> catholiques, </pers.coll> qui n'a pu
paraître en notre édition d' <time.date.rel> <name> hier </name> <time-modifier> soir. </time-modifier> </time.date\
     .rel>
Ce document trace avec une grande
autorité le programme qui fait le fond
des revendications de la <org.ent> <kind> Ligue </kind> de
l' <name> "Ave Maria" </name> </org.ent> et dont il faut pour-
suivre, ajoute l' <func.ind> <qualifier> Eminent </qualifier> <kind> prélat, </kind> </func.ind> la réa-
lisation "avec calme, énergie et persé-

00212633/PAG_1_TB000013.png
vérance".

00212633/PAG_1_TB000014.png
i Le <func.ind> <kind> cardinal </kind> </func.ind> termine en disant qu'il
faut faire des oeuvres et prier.
Nous recevons ce document avec
d'autant plus de joie qu'il devient une
sorte de consécration de nos efforts.
```

which is the case of OCR output. The main limit is that the correct / incorrect decision is binary. Experience has shown that errors on type selection and errors on boundaries are two separate issues, even though somewhat interdependent. The alternative is to go for an *error counting* method where errors are enumerated and costs are associated. That method produces the *Slot Error Rate* (Makhoul et al., 1999), which we used in our evaluation, with a cost of 0.5 for type errors and 0.5 for boundary errors. The detail of the measurement methodology, including how to handle structured entities, is described in (Galibert et al., 2011).

4.3. Results and discussion

Three systems participated in the evaluation. We computed raw results which are given in Table 2.

	Precision	Recall	F-measure	SER
System 1	68.9%	61.8%	65.2%	44.2%
System 2	68.6%	49.6%	57.6%	50.0%
System 3	55.2%	46.4%	50.4%	60.3%

Table 2: Evaluation results for the three participants.

The two best performing systems are mostly stochastic, with heavy use of CRFs and probabilitistic parsing, while the third system is essentially linguistic (resource- and rule-based). It is interesting to note that in clean contexts the third system tends to have much better results, whereas its ranking in the present evaluation shows in part the difficulty of handling OCR errors, producing lots of broken words and punctuations, with syntactically and semantically deep approaches.

An analysis of the missed entities showed that even statistical systems were seldom able to extract entities with 'broken characters' unless the context was very strong, as was the case in "le 17 mars 183t," where the statistical systems did recognize "183t" as the year while the linguistic system did not. In general nevertheless, they seem to have been less sensitive to 'broken context'.

Finally, system results are indeed partly explained by the difficulty of the task itself, OCR or not: on the one hand, old texts written in a somewhat dated French where topics, location and person names are different from what can be found in current news; and on the other hand, novel annotation guidelines. That alone made the task quite challenging.

5. Conclusion and perspectives

Within the framework of the Quaero program, we launched an evaluation campaign in the domain of named entity recognition. This evaluation was performed on a corpus composed of OCRed press archives from French newspapers of December 1890. We first presented the methodology we used to select the text blocks to be annotated. Then, we detailed the annotation process based upon the definition we proposed previously for extended named entities. Finally, we presented the evaluation performed on the annotated old press corpus and the metrics we used to compute the final scores. Three systems, mainly based on statistical methods, participated in this evaluation. The computed slot error rate ranged from 44.2% to 60.3%.

While this evaluation did not lead very high results, the approaches used by the participants to deal with the characteristics of both the corpus and the annotation schema allowed us to validate the definition we proposed for extended named entities.

In future work, we plan to organize other evaluations, especially open ones. As a way to avoid the scarcity of available corpora annotated with named entities, we are also taking steps to make the corpus available to the scientific community.

Acknowledgments

This work was realized as part of the Quaero Programme, funded by Oseo, French State agency for innovation. We thank all the annotators *(Jérémy, Matthieu, Orane, and Raoum)* of the ELDA company; they worked seriously and with professionalism.

6. References

Eckhard Bick. 2004. A Named Entity recognizer for Danish. In *Proc. of LREC*, Lisbon, Portugal. ELRA.

Kate Byrne. 2007. Nested Named Entity Recognition in Historical Archive Text. In *Proceedings of the first IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, California.

Richard Tobin Claire Grover, Sharon Givon and Julian Ball. 2008. Named Entity Recognition for Digitised Historical Texts. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/.

Sam Coates-Stephens. 1992. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities*, 26:441–456.

Gregory Crane and Alison Jones. 2006. The challenge of Virginia Banks: An evaluation of named entity analysis in a 19th-century newspaper collection. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 31–40, New York, NY, USA. ACM.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation. In *Proc. of LREC*, pages 837–840. ELRA.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *Proc. of COLING*, pages 1–7. Association for Computational Linguistics.

Michael Fleischman. 2001. Automated Subcategorization of Named Entities. In *Proc. of the ACL 2001 Student Research Workshop*, pages 25–30.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. 2010. Named and Specific Entity Detection in Varied Data: The Quaero Named Entity Baseline Evaluation. In *Proc. of LREC*, Valletta, Malta. ELRA.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2011. Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In *Proc. of IJCNLP*, Chiang Mai, Thailand.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proc. of InterSpeech*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proc. of COLING*, pages 466–471, Copenhagen, Denmark.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proc. of the Fifth Linguistic Annotation Workshop (LAW-V)*, Portland, OR. Association for Computational Linguistics.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, pages 249–252.

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: speech and OCR. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 316–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum, 2011. *Entités Nommées Structurées : guide d'annotation Quaero*. LIMSI–CNRS, Orsay, France. `http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf`.

Satoshi Sekine. 2004. Definition, dictionaries and tagger of extended named entity hierarchy. In *Proc. of LREC*, Lisbon, Portugal. ELRA.