Diachronic Evaluation of NER Systems on Old Newspapers

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, Frédéric Kaplan
Digital Humanities Laboratory (DHLAB)
Swiss Federal Institute of Technology in Lausanne (EPFL)
CDH, INN 116, Station 14, Lausanne, Switzerland
name.surname@epfl.ch

Abstract

In recent years, many cultural institutions have engaged in large-scale newspaper digitization projects and large amounts of historical texts are being acquired (via transcription or OCRization). Beyond document preservation, the next step consists in providing an enhanced access to the content of these digital resources. In this regard, the processing of units which act as referential anchors, namely named entities (NE), is of particular importance. Yet, the application of standard NE tools to historical texts faces several challenges and performances are often not as good as on contemporary documents. This paper investigates the performances of different NE recognition tools applied on old newspapers by conducting a diachronic evaluation over 7 time-series taken from the archives of Swiss newspaper *Le Temps*.

1 Introduction

Recognition and identification of real-world entities is at the core of most text mining applications. As a matter of fact, referential units such as names of persons, locations and organizations underlie the semantics of texts and guide their interpretation. Since the seminal MUC shared-task (Grishman and Sundheim, 1996), named entity-related tasks have undergone major evolutions, from entity recognition and classification to entity disambiguation and linking (Nadeau and Sekine, 2007; Rao et al., 2013). Besides the general domain of well-written news-wire data, NE processing is also applied on specific domains, particularly biomedical (Kim et al., 2003), and on more noisy inputs such as speech transcriptions and tweets (Galibert et al., 2014; Ritter et al., 2011). More recently, NE processing has also been called upon to contribute to the domain of digital humanities, where massive digitization of historical documents is producing huge amounts of texts.

In the last few years, many cultural institutions have indeed engaged in large-scale digitization projects (Gerhard and van den Heuvel, 2015), some with a general scope, e.g. Europeana[1] or CultureSampo[2], others focusing on specific resources such as historical newspapers, e.g. Europeana Newspaper[3] (Neudecker and Antonacopoulos, 2016) or the National Digital Newspaper Program[4]. Millions of images are being acquired and, when it comes to text, their content is transcribed, either manually via dedicated interfaces, or automatically via Optical Character Recognition (OCR). If this represents a major step forward in terms of preservation and document accessibility, much remains to do in order to provide an extensive and sophisticated access to the *content* of digital resources. In this regard, information extraction techniques, particularly NE extraction and linking, can certainly be regarded as among the first steps.

Historical documents, however, pose many challenges for language technologies (Sporleder, 2010). Due to the acquisition process and/or the conservation state, input texts can be extremely noisy. Next, language(s) of earlier stage(s) may feature old vocabulary and turns of phrases and, in the case of NE extraction, can contain entities for which adequate linguistic resources and knowledge bases are missing (Ehrmann et al., 2016). Finally, as demonstrated by Vilain et al. (2007), the transfer of NE tools from one domain to another is not straightforward and performances of NE tools, initially developed for homogeneous texts of the

[1] http://www.europeana.eu/portal/about.html

[2] http://www.kulttuurisampo.fi/about.shtml?lang=en

[3] http://www.europeana-newspapers.eu/

[4] https://www.loc.gov/ndnp/

immediate past, are very likely to be affected by these phenomena.

Named entity processing tools are particularly requested in the context of historical newspapers, where historians wish to discover, among others, the “5 W’s”: *who did what when and where with whom*. In this paper, we experiment with the application of prototypical NE recognition and classification (NERC) approaches on a newspaper digital archive. More specifically, we are interested in investigating whether the performances of NE tools degrades when going back in time. To this end, we apply 4 NER systems on 7 document time-series (1804 to 1981) from the archives of French speaking Swiss newspaper *Le Temps*.

The remainder of the paper is organised as follows. Section 2 presents the main challenges of NE processing on historical text and discusses how they were tackled in related work. Next, section 3 describes our experimental settings, with the presentation of the source (section 3.1), the evaluation data set (section 3.2) and the systems (section 3.3). Section 4 details the results and provides an error analysis and, finally, section 5 concludes and considers future work.

2 Named Entity Processing for Cultural Heritage domains

Along with the increasing demand for language technologies support for cultural heritage domains, recent years have seen a surge in research on NE processing for historical texts. Work in this domain can be divided according to the nature of the texts which is dealt with (e.g. museum record metadata, administrative documents, genealogical data, newspapers), according to the written modality (handwritten or typeset), and according to the targeted task (NE recognition and classification, entity linking, or both). Most experiments follow one of the two following strategies: application and/or tuning of an already existing system (available in-house or publicly released, e.g. Stanford NER[5]), or use of NE processing web-services. Overall, existing work concerns a wide variety of texts covering different historical periods (from 16^{th} to 20^{th} c.), focus on different domains and use different typologies. This great variety demonstrates how many and varied the needs – and the challenges – are, but makes performance comparison difficult, not to say impossible.

Compared to the standard analysis of present-time English, very often news, the application of NE tools on historical texts faces news challenges, which can be defined as follows: (i) noisy input texts, (ii) lack of coverage in linguistic resources and knowledge bases, and (iii) dynamics of language. This section briefly elaborates on these challenges.

2.1 Noisy input texts

Texts acquired from digitized historical material can be extremely noisy. Errors can be caused either by the original source, e.g. degraded material or non standardized language, or from processing effects, e.g. poor OCR quality. They do not resemble tweet misspellings or speech transcription hesitations, problems for which adapted approaches have already been devised (Ritter et al., 2011; Parada et al., 2011).

Language variation was successfully tackled by Borin et al. (2007), who tuned an existing rule-based system with a name similarity calculation mechanism. Working on Swedish literary classics from the 19^{th}c., they were able to recognize entities belonging to 8 categories with a F-measure of 92.8%.

In some contexts, OCR errors have been handled positively, e.g. as part of the French *Quaero* project[6]. First, a comparative study of structured NE manual annotation in broadcast news vs. 19^{th}c. historical newspapers (*Le Temps*, *La Croix* and *Le Figaro* of December 1890) has been conducted, showing that OCR noise requires some guideline adaptations (Rosset et al., 2012). Three systems were subsequently evaluated on the annotated data with a F-measure ranging from 57.6% to 65.2% (Galibert et al., 2011a). Later on, Dinarelli and Rosset (2012) implemented several OCR correction strategies on this material, leading to a reduction of SER (Slot Error Rate, explained hereafter) of 8 points.

However, it appears sometimes that not even dedicated manual efforts seem to improve the quality of the recognition for historical data. Rodriquez et al. (2012) compared the performances of four NER system (Stanford, OpenNLP, AlchemyAPI and OpenCalais) on two data sets related to WWII: individual Holocaust survivor testimonies from the Wiener Library of London and letters

[5] http://nlp.stanford.edu/software/CRF-NER.shtml

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

of soldiers from King's College archive. Performances are evaluated against a (small) gold standard comprising person, location and organisation names. Results on OCRed data are between 47% and 54% F-measure for the testimonies (Stanford being the most accurate), and between 32% and 36% for letters (OpenCalais performing best). When applied on manually corrected OCR, tools performed better, but not significantly. Other major identified sources of errors are different ways of naming and metonymy phenomena (esp. war ships named after people) and lack of knowledge of the systems (esp. for organisations).

2.2 Poor resource coverage

Many NE tools rely, at least in part, on existing linguistic resources and knowledge-bases, such as Wikipedia/DBpedia. However, the coverage of any knowledge base is at best uneven when going back in time. Three different phenomena are likely to impact on the lack of proper coverage in knowledge-bases: mentions of minor or not well-known entities, entities that changed name over time, and names that were used for different entities over time (ambiguity).

The general poor performance of knowledge-based systems was highlighted for example by Hooland et al. (2015). They aimed at indexing the descriptive fields of records from the Cooper Hewitt museum of New York. To this end, they developed an OpenRefine NER extension based on multiple NER web-services (AlchemyAPI, DBpedia Spotlight and Zemanta), giving the possibility for data curators to automatically annotate and link entities within records. Evaluation was done against a manually built gold standard with 4 categories, with a F-measure ranging from 10 to 60% and a low recall for all systems.

Nevertheless, others found it possible to rely on knowledge bases in order to enrich them. As an example, Huet et al. (2013) explored how to mine history from *Le Monde* French newspapers (issues between 1944-1986) by linking entities occurring in articles to YAGO referents. Entities are broadly defined (we assume all entity types of YAGO) and their recognition is done via a look-up procedure, with a Precision of 86.8% and a Recall of 77.1%.

2.3 Dynamics of language

The last source of errors, and to the best of our knowledge the least explored by research up to date, relates to the dynamics of language. Most projects dealing with historical textual data cannot assume that similar rules and conventions for the use of written language applied at all times. Some previous studies showed how older data might be more problematic. Grover et al. (2008) focused on British parliamentary proceeding from the end of the 17^{th} and the beginning of 19^{th} centuries. OCRed documents are given as input to an in-house rule-based system in order to extract person and place names. The overall performance is evaluated against a gold standard of ca. 6000 person and 3600 place names, with an F-measure of about 70% for both periods. Results are comparable for person names, but the earliest period has significantly worse performance for locations.

In order to compensate for lack of dedicated studies on the problem of language change over time and to better understand NE recognition performances on historical texts, we conducted a diachronic evaluation of different NER tools over 200 years of historical newspapers. This work is in line with both (Rodriquez et al., 2012) and (Hooland et al., 2015) who applied different NE tools on historical texts, and (Galibert et al., 2010) and (Rosset et al., 2012) who explored NE annotation on French old newspapers. However, our approach features web-based NE annotation tools – never evaluated on newspapers to the best of our knowledge – and considers time series data sets. Those times series are derived from the archives of *Le Temps* newspaper, established in the French speaking part of the Swiss Confederation.

3 Experimental setting

3.1 *Le Temps* digital archive

The Swiss newspaper *Le Temps* originates from the merger of *La Gazette de Lausanne* (GDL), *Le Journal de Genève* (JDG) and *Le Nouveau Quotidien* in 1998. Born in 1798, 1826 and 1991 respectively, these three publications compose the digital archive of *Le Temps*, which was acquired in 2008 via optical character recognition (OCR) and layout detection. Together, the *Gazette de Lausanne* and the *Journal de Genève* comprise about 1 million pages and 4 million articles spanning 200 hundred years of Swiss and international history. Taking a linguistic, historical or sociological view point, motivations to explore this collection of past events and society are manifold (Bingham, 2010), and named entity recognition can in this regard be of great assistance.

	# words		# pers		# loc		# entities	
	GDL	JDG	GDL	JDG	GDL	JDG	GDL	JDG
1804	33,773	-	417	-	990	-	1,407	-
1826	33,353	14,074	471	184	946	151	1,417	335
1841	40,784	5,558	553	70	1,137	55	1,690	125
1881	55,751	12,360	950	227	912	280	1,862	507
1921	20,117	3,587	377	47	572	136	949	183
1961	23,332	8,301	529	115	556	149	1,085	264
1981	17,759	3,672	258	79	363	56	621	135
TOTAL	299,212	65,139	3,555	722	5,476	827	9,031	1,549

Table 1: Data set statistics.

3.2 Data set

We randomly selected 40 article files from GDL and 10 from JDG for the years 1804, 1826, 1841, 1881, 1921, 1961 and 1981[7]. The choice of these years was not motivated by any specific historical events but to ensure even coverage of the period. Article files were built by parsing the XML output of the OCR system, that is to say by re-building the text from the xml-tagged token singletons, and by assembling different text blocks belonging to the same article.

The selected files were annotated according to the *Quareo* guidelines (Rosset and Grouin, 2011), which have already been used for the annotation of French historical newspapers. With this choice the present data will therefore contribute to the constitution of a larger and diversified set of NE-annotated historical newspaper corpora and, on the long run, ensure performance comparison. *Quareo* typology is both hierarchical and compositional with, on the one hand, 7 entity types and 32 sub-types which categorize entities and, on the other, 24 entity components which specify the various elements making up the entities. For the present annotation task we did not considered components and targeted exclusively Person and Location entity types, with their relative *Quaero* subtypes (*pers.ind*, *pers.coll*, *loc.adm.reg*, *loc.admin.nat*, etc.).

Manual annotation was carried out from scratch by the authors (two native French speakers, and one fluent in French) using the brat rapid annotation tool (Stenetorp et al., 2012). As noticed by Rosset et al. (2012), annotation of old texts is possible but not straightforward. Annotation was done without looking at the image of the articles, that is to say relying on the OCRed text only. In this regard, decision was made to annotate entity mentions containing OCR noise as far as the annotator could recognize and identify them (e.g. *Constap. iipopjle*). The reason why we included noisy entities is because "OCR name variants" can legitimately be recognized and can be useful in an information retrieval or text mining application context. A bot or an information seeking person would indeed certainly be interested in retrieving docs in which the original text was referring to a certain entity, whatever its OCR transcription. As for historically moving entities, annotation was done according to their more recent status (e.g. *Malte* annotated as *loc.adm.nat*). Also, it should be noted that nested entities are annotated, e.g. in *Bern University*, *Bern* is annotated as Location and *Bern University* should be – but we do not consider this type at the moment – annotated as Organization. Finally, according to *Quaero* guidelines, titles such as *M.*, *Mme*, *Mlle* are part of person names, whereas functions such as *prime minister* are not.

In order to estimate the quality of the annotation, agreement rate between the 3 annotators has been computed over 3 documents of GDL from 1826, 1921 and 1981. Fleiss coefficient (Fleiss, 1971) with boundary fuzzyness on fine and coarse-grained types corresponds to 0.88 and 0.95 resp., which can be considered as satisfactory.

Table 1 shows the overall statistics of the annotated texts. Among the two newspapers, GDL is the biggest corpus; it gathers 280 articles with 3555 person and 5476 location names, for a total of about 300k words. 1881 is the year with the most entities, 1921 with the less. JDG is more reduced, with only 60 articles, 722 persons, 827 locations and about 65k words. In both corpora, the overall number of entities first increases and then decreases. This could be connected with the evolution of articles' length, getting longer during the 19^{th} c., and shorter during the 20^{th} c.

3.3 Systems

Four systems were included in our study. With the primary condition of having parsing capacities for French language, the selected tools represent major approaches for NERC: symbolic system with ExPRESS, supervised machine learning with mXS[8] and proprietary web services offering NER functionalities with AlchemyAPI[9] and Dan-

[7] JDG starts in 1826 only.

[8] `https://github.com/eldams/mXS`

[9] `http://www.alchemyapi.com/products/alchemylanguage/entity-extraction`

delionAPI[10,11]. In all experiments systems have been applied out of the box without any adaptation.

Rule-based system This NERC system consists of a set of manually curated language-independent rules that make use of language specific lexicons encoding information about entity names and trigger words. Defined via the extraction pattern engine ExPRESS (Piskorski, 2007), rules are modelled as a cascade of finite-state grammars where units are processed in increasing order of complexity. Apart from a light pre-processing including tokenization and sentence splitting, no morphological analysis nor POS-tagging is required. In concrete terms, NE rules focus on typical patterns of person, location and organisation names, e.g. an adjective (*former*) followed by a function name (*President of the Confederation*), a first (*Ruth*) and a last (*Dreifuss*) name. Besides modifiers (*famous*) and function names (*minister*), trigger words cover professions (*guitarist, football player*), expression indicating age (*42 years-old*), demonyms and markers of religion or ethnical groups (*Italian, Genevan, Bambara, Muslim*), and more. This system is derived from the multilingual NER framework developed in the context of the *Europe Media Monitor* (EMM) (Steinberger et al., 2009), from where originates the entity resource JRC-Names (Steinberger et al., 2011; Ehrmann et al., 2016). This system is tuned to recognize at least one mention per documents and is therefore better at precision than recall. In this work only the French grammar is considered.

mXS is a supervised machine learning system which learns extraction patterns for named entities. The specificity of mXS (Nouvel et al., 2014) is that it tries to detect separately the left and right boundaries of entities, a strategy particularly useful with noisy texts such as speech transcriptions where boundary markers differ due to hesitations and disfluencies. Using data mining techniques, the model first learns extraction patterns, before applying filters and a Maximum Entropy classifier over the patterns. Its performance has been evaluated against the ETAPE French corpora of speech transcriptions (Gravier et al., 2012) with a Precision of 79.8% and a Recall of 64.9%.

AlchemyAPI The AlchemyAPI is a hybrid system which combines supervised classification and rules based on textual cues to perform NERC and disambiguation. The backbone knowledge graph is proprietary; it includes all main open KBs and entities are disambiguated towards, among others, DBpedia, Freebase, GeoNames, Census and OpenCyc.

DandelionAPI Dandelion is based on a knowledge graph built from several repositories and mostly composed of places, events, organisations and people (Parmesan et al., 2014). The backbone of this knowledge graph is DBpedia, whose textual content and internal entity relations are used to perform NERC and disambiguation. In this work both Alchemy and Dandelion are used for their entity recognition and classification capacities only.

4 Evaluation

4.1 Metrics

System performances are evaluated in terms of precision and recall for each time period and in terms of their aggregation over all entities across all documents, that is to say Micro-Average precision and recall (MAP/R), for the whole period. In both cases the harmonic mean F-measure (F1) is also reported.

As demonstrated by Makhoul et al. (1999), if these measures are good at evaluating what is correct (or not), they however do not fully nor truly account for errors, especially the F-measure. As a consequence, we additionally consider the Slot Error Rate (SER), a measure analogous to the Word Error Rate in speech recognition, computed as follows:

$$SER = \frac{D + I + STB + 0.5 \times (ST + SB)}{R} \qquad (1)$$

where D corresponds to the number of *Deletions* (false negatives), I to the number of *Insertions* (false positives), ST to the number of *Type Substitutions*, SB to the number of *Boundary Substitutions*, STB to the number of *Type and Boundary Substitutions* (i.e. items with incorrect type and boundaries but having a common component with an item of the reference) and R to the total number of reference entities. The adopted weighting scheme is similar as in (Galibert et al., 2011a) and gives less importance to type or boundary substitutions. Contrarily to the previous measures, SER

[10] https://dandelion.eu/

[11] Proprietary web-services were used during May 2016. We thank both IBM (Alchemy) and Spazio Dati (Dandelion) for willingly providing free API access for the purpose of this research.

	Dandelion			Alchemy			Rule-based			mXS		
	P	R	F	P	R	F	P	R	F	P	R	F
1804	20.4	11.0	14.3	53.7	27.8	36.7	62.4	12.7	21.1	28.3	18.9	22.7
1826	20.1	9.6	12.9	46.3	31.0	37.2	61.9	14.9	24.0	26.0	22.1	23.9
1841	24.4	11.6	15.7	60.3	33.3	42.9	69.1	11.8	20.1	25.7	17.2	20.6
1881	26.0	8.7	13.1	73.7	40.4	52.2	67.2	14.0	23.2	38.8	26.5	31.5
1921	38.1	13.5	20.0	72.0	41.6	52.8	69.6	23.1	34.7	32.2	23.3	27.1
1961	39.0	22.1	28.2	73.3	51.4	60.4	67.5	25.5	37.0	41.6	27.8	33.3
1981	29.9	30.6	30.3	**75.9**	**56.2**	64.6	72.8	41.5	52.8	30.6	31.8	31.2
All years	28.1	13.6	18.4	65.7	**39.5**	**49.3**	**67.6**	18.3	28.8	32.7	23.8	27.6
Baseline	52.8	34.3	41.6	**86.7**	55.6	67.7	86.3	39.7	54.4	77.3	**72.8**	**75.0**

Table 2: Precision, Recall and F-measure for *Person* on GDL corpus, plus Baseline on Quaero corpus.

	Dandelion			Alchemy			Rule-based			mXS		
	P	R	F	P	R	F	P	R	F	P	R	F
1804	63.4	64.3	63.9	63.7	28.2	39.1	90.1	43.0	58.2	71.4	32.2	44.4
1826	60.8	64.1	62.4	59.0	25.7	35.8	85.0	45.6	59.3	69.0	33.8	45.4
1841	70.6	74.0	72.3	55.4	28.1	37.3	**91.1**	51.2	65.6	70.6	35.7	47.5
1881	51.4	68.3	58.7	55.0	33.6	41.7	77.2	53.7	63.3	62.0	38.0	47.1
1921	65.2	**75.3**	69.9	53.2	28.7	37.3	87.0	50.2	63.6	63.3	30.8	41.4
1961	54.2	69.2	60.8	60.9	27.7	38.1	82.3	34.2	48.3	68.0	30.6	42.2
1981	52.2	67.2	58.8	50.2	29.5	37.2	72.5	39.9	51.5	62.0	34.2	44.0
All years	60.4	**68.8**	**64.3**	57.0	28.7	38.2	**84.6**	46.6	60.1	67.1	34.0	45.1
Baseline	57.5	**77.7**	66.1	50.6	35.7	41.8	84.7	66.0	74.2	**85.2**	68.8	**76.1**

Table 3: Precision, Recall and F-measure for *Location* on GDL corpus, plus Baseline on Quaero corpus.

is not a figure of merit but of error, therefore the lower its value the better the performance of the system. Under high error conditions, SER can be greater than 1.

4.2 Results and Error Analysis

The discussion focuses on Tables 2 and 3 which show results for the four systems in terms of precision, recall and F-measure for the GDL data set. Tables 4 and 5 report on the same measures but with a "fuzzy" setting where boundary mistakes are accepted. Given that all systems do not follow the same annotation conventions than the one we adopted, this tolerant evaluation scheme allows for a better comparison of systems. Annotation differences include insertion or not of titles and functions in person names (e.g. `<pers>` *chancellor Adenauer* `</pers>` vs. *chancellor* `<pers>` *Adenauer* `</pers>`), and of specifiers in location names (`<pers>` *district of Nyon* `</pers>` vs. *district of* `<pers>` *Nyon*`</pers>`). Regarding titles and functions, recall that Quaero guidelines include the former but exclude the latter (cf. section 3.2); in this regard, mXS and Dandelion are penalized for they exclude titles, Alchemy for it includes functions. As for locations, *Quaero* ask for the annotation of specifiers; all systems exclude them and are penalized in the same way. Finally, Figure 1 render the same measures in a graphical manner and Tables 6 and 7 present the Slot Error Rates for the Person type. The comparison with JDG is omitted for brevity as it largely confirms those from GDL.

Baseline As we wish to assess the performance gaps of NERC tools between present and historical texts (in this case newspapers), we compute a baseline against one of the few recent gold standard for French: the test data of the *Quaero* Broadcast News evaluation campaign (Galibert et al., 2011b). It is composed of speech transcriptions of radio and TV broadcasts from the year 2010; 1386 entities of type Person and 747 of type Location[12] are annotated according to the *Quaero* annotation conventions. Baseline figures are shown in last rows of Tables 2, 3, 4 and 5. In both settings and for both types, mXS (trained on speech data) performs best, with F-measures of 75% (Per-

[12] We did not consider all annotations but only the ones corresponding to our data sets.

	Dandelion			Alchemy			Rule-based			mXS		
	P	R	F	P	R	F	P	R	F	P	R	F
1804	37.3	20.1	26.2	70.4	36.5	48.0	88.2	18.0	29.9	45.9	30.7	36.8
1826	43.3	20.6	27.9	70.2	46.9	56.2	92.0	22.1	35.6	49.8	42.3	45.7
1841	53.4	25.3	34.4	76.1	42.0	54.1	**97.9**	16.6	28.4	40.9	27.3	32.8
1881	48.6	16.3	24.4	87.7	48.1	62.1	96.0	20.0	33.1	59.9	40.9	48.7
1921	64.2	22.8	33.7	89.9	52.0	65.9	92.0	30.5	45.8	53.1	38.5	44.6
1961	55.0	31.2	39.8	89.2	62.6	73.6	94.0	35.5	51.6	58.1	38.8	46.5
1981	41.3	42.2	41.8	85.9	**63.6**	73.1	95.2	54.3	69.1	44.4	46.1	45.2
All years	48.4	23.5	31.6	82.0	**49.3**	**61.6**	**94.0**	25.4	40.0	51.6	37.6	43.5
Baseline	59.5	38.6	46.8	96.5	61.8	75.4	**97.3**	44.7	61.3	86.9	**81.8**	**84.3**

Table 4: *Fuzzy* Precision, Recall and F-measure for the type *Person* on GDL corpus.

	Dandelion			Alchemy			Rule-based			mXS		
	P	R	F	P	R	F	P	R	F	P	R	F
1804	67.5	68.5	68.0	**94.5**	41.8	58.0	92.6	44.2	59.9	76.7	34.6	47.7
1826	68.8	72.4	70.5	93.2	40.6	56.6	89.5	48.0	62.5	75.0	36.8	49.4
1841	75.1	78.8	76.9	92.4	46.8	62.1	92.0	51.8	66.3	74.4	37.7	50.0
1881	55.5	73.7	63.3	76.3	46.5	57.8	80.2	55.8	65.8	68.9	42.3	52.4
1921	68.4	**79.0**	73.3	89.9	48.4	63.0	87.9	50.7	64.3	70.1	34.1	45.9
1961	59.2	75.5	66.4	85.8	39.0	53.6	86.1	35.8	50.6	71.2	32.0	44.2
1981	58.7	75.5	66.0	76.1	44.6	56.3	83.5	46.0	59.3	69.5	38.3	49.4
All years	65.3	**74.4**	**69.6**	87.4	44.0	58.6	**87.7**	48.3	62.3	72.7	36.8	48.9
Baseline	63.9	**86.3**	73.4	75.9	53.5	62.7	86.2	67.2	75.5	84.3	71.4	**79.1**

Table 5: *Fuzzy* Precision, Recall and F-measure for the type *Location* on GDL corpus.

son) and 76.1% (Location) in normal setting and of 84.3% and 79.1% in fuzzy setting. Regarding Person, Alchemy and the rule-based (RB) systems score high in precision whereas recall is lower, particularly for RB. Dandelion is overall better than Alchemy for Location, but performs equally than RB on this type.

General observations In terms of precision, performances over all years ranges from 28.1% to 67.6% for the type Person and from 57% to 84.6% for the type Location (cf. Tables 2 and 3). In terms of recall, performances reach values from 13.6% to 39.5% (Person) and from 28.7% to 68.8% (Location). Best F-measures correspond to 49.3% for Person and 64.3% for Location. When considering the fuzzy scheme (cf. Tables 4 and 5), performances are better, particularly for Person's precision and recall which show success rates at 94% and 49.3%, respectively. Location's performances increase as well but not that greatly. In this setting, best F-measures reach 61.6% and 69.6% for Person and Location respectively. High slot error rates echo these figures, with 0.63 for Person and 0.58 for Location at minima (cf. Tables 6 and 7 for Person; Location tables are omitted).

Not surprisingly, these results do not compare with the mid-90s F-score achieved by the MUC systems and are below the usual performances on news genre; they are however in line with the figures obtained on historical newspapers in (Galibert et al., 2010; Galibert et al., 2011a). Overall, the situation is better better for Location than for Person in terms of both precision and recall, and performances show important disparities between systems.

Compared to the baseline, all systems show degraded performances. Overall, losses are more important for Person than for Location and are different among systems. mXS is the most affected, with F-measure downgraded by 40.8 points on Person and 30.2 on Location (fuzzy setting). Alchemy and RB have important losses regarding Person, Alchemy rather on precision (−14.5 points on fuzzy), RB rather on recall (−19.3). Dandelion is mainly affected on the Person type, generally, and on Location, for recall only.

Considering general performances on the historical corpus, the rule-based system stands on the podium during the first half of the period in terms of Person precision, before being overtaken by

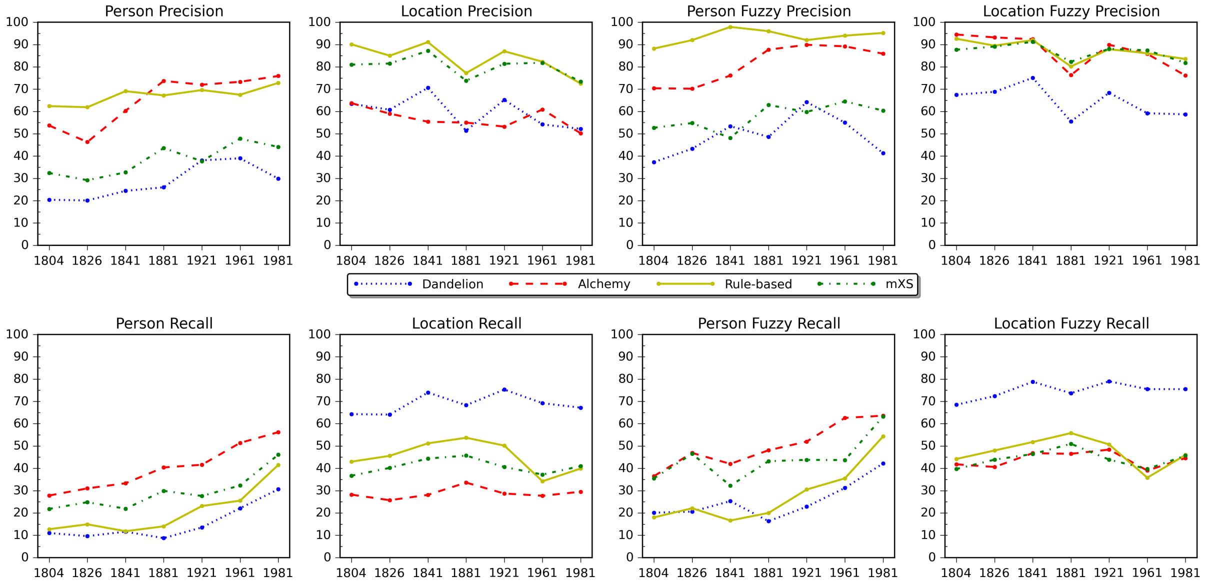

Figure 1: Precision and Recall plots for all systems, with normal and fuzzy settings.

Alchemy for the second half. It however stays the first system for Location precision, very closely followed by mXS. With respect to recall, Alchemy and Dandelion show opposite and reversed performances: Alchemy is best for Person but worst for Location, and the contrary for Dandelion. The same holds true with the fuzzy setting.

Time-based observations For both types evolution of system's precision over time is quite irregular, with several ups and downs for all systems, except RB and Alchemy which are slightly more stable for Person and Location respectively (cf. Figure 1). Similar trends can be observed under the fuzzy scheme. Contrary to what could have been expected, precision do not show clear increase over time, since the situation kind of improves for Person, and even degrades for Location. On the opposite, recall show less variability over time, with a slight but regular increase for Person towards the year 1981, and a more stable situation for Location. We may conclude that the way location names are introduced in texts is more stable than for person names, and that the contribution of knowledge bases (or gazetteers in the case of RB and mXS) is in this case more profitable.

System-based observations Considering the different systems, we observe important performances discrepancies in both absolute terms and time-related trends; the ease and the difficulties are not the same for all systems. The most stable systems over the years are RB for Person's precision and mXS for Location's recall. In terms of overall precision, Alchemy and RB are good for Persons, while mXS and RB systems are efficient for Locations. Person's top precision is reached by Alchemy in normal setting and by RB in fuzzy setting; Location's top one by RB (normal setting) and Alchemy (fuzzy setting). As for recall, Alchemy is the best for Person, Dandelion for Location, while mXS shows a better balance over both types.

Tables 6 and 7 detail the various types of errors in terms of SER variables (on Person type only; however, we also report figures on Location hereafter). For both types Dandelion has the highest number of *Insertions*; their evolution through time is irregular for Person, while they regularly decrease for Location. For all systems the number of *Deletions* evolves quite irregularly, but is lower at the end of the period than at the beginning. Systems who deleted most entities are Dandelion and RB for Persons, and Alchemy and mXS for Locations. Dandelion and RB do not confuse Person types, but can do mistakes for Location. Alchemy and mXS often mistaken Person for Location, but less Location for Person.

4.3 Discussion

This diachronic analysis allowed us to peek under the hood of different NE tools challenged with texts from historical newspapers. Despite the fact that this is a first evaluation, some trends emerge. First, performances degrade compared to

	Dandelion						Alchemy					
	I	*D*	*ST*	*SB*	*STB*	SER	*I*	*D*	*ST*	*SB*	*STB*	SER
1804	132	309	1	38	8	1.12	40	262	18	36	6	0.8
1826	123	342	3	52	1	1.05	43	241	47	75	4	0.74
1841	116	374	3	76	4	0.96	36	307	33	48	5	0.7
1881	150	753	2	72	14	1	33	464	29	73	3	0.58
1921	45	278	3	35	0	**0.91**	11	180	10	39	1	0.57
1961	122	351	10	48	3	0.95	29	191	10	59	20	0.52
1981	148	139	1	30	6	1.2	16	91	10	19	1	**0.47**
All years	836	2546	23	351	36	1.01	208	1736	157	349	40	**0.63**

Table 6: Alchemy and Dandelion SER results for type *Person* on GDL corpus.

	Rule-based						mXS					
	I	*D*	*ST*	*SB*	*STB*	SER	*I*	*D*	*ST*	*SB*	*STB*	SER
1804	2	325	0	22	8	0.83	110	270	26	49	17	1.04
1826	7	345	2	34	0	0.79	155	268	41	95	5	1.05
1841	1	438	1	27	2	0.82	152	400	61	56	11	1.12
1881	6	708	1	57	4	0.79	216	562	43	137	3	0.92
1921	4	252	1	28	6	0.73	96	229	19	57	14	1
1961	12	329	0	53	2	0.7	122	317	24	58	2	**0.91**
1981	7	112	0	33	0	**0.53**	118	138	31	37	0	1.12
All years	39	2509	5	254	22	0.76	969	2184	245	489	52	1

Table 7: Rule-based and mXS SER results for type *Person* on GDL corpus.

the adopted baseline and are lower than those observed during traditional NE evaluation campaigns such as MUC or CoNNL. However, they are in line with other work on historical newspapers.

Next, results show more irregularities over time than expected, as well as strong disparities between systems. Nevertheless, the historical trend for Location recall confirms the intuition that the more recent the texts, the more entities we can recognize. This suggest that the lexical coverage of gazetteers and/or knowledge bases (which constitutes the backbone of some systems) is lower when going back in time. Then, the significant performance drop on earlier years (especially for recall) might be due to a lower OCR quality and to text variability. We tend to discard a strong impact of language variability issues afterwards, since newspapers were commonly proofread. The same applies to OCR impact, for which an evaluation campaign is ongoing.

Finally, performances over historical newspapers vary depending on entity types. Contrarily to Persons, Location names can be expected to be mentioned in a more stable way over time; this is confirmed by higher performances on this type, especially in terms of recall and for systems relying on knowledge bases.

Regarding the best strategy to follow in order to adopt an NE tool to process historical newspapers, this analysis shows that no clear-cut solution exists: all tools have strengths and weaknesses either over time, or over specific types of NEs, or over recall and/or precision optimization. The best solution might therefore be to make a diachronic evaluation and then select or combine the best tools for a given period, a given type of entity and a given preferred application scenario.

5 Conclusion and Future work

We presented a diachronic evaluation of 4 NERC tools applied to 7 time-series from Swiss newspaper archive *Le Temps*. The evaluation spans almost 200 years and allows to understand better the behaviour of NE tools on historical data. Performances are overall lower than on contemporary texts and, interestingly, the intuition that they degrade when going back in time is only partially validated: it holds true for the Location type but not for Person.

Many directions remain open as future work. We intend to evaluate the impact of OCR errors, to expand our NE set to the full *Quaero* typology, and to consider others historical data sets. Such developments will lay the ground for advanced text mining over *Le Temps* corpus and, more generally, over historical newspapers.

References

A. Bingham. 2010. 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2):225–231.

L. Borin, D. Kokkinakis, and L-J. Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century swedish literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaT-eCH 2007)*, pages 1–8.

M. Dinarelli and S. Rosset. 2012. Tree-Structured Named Entity Recognition on OCR Data: Analysis, Processing and Results. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the* 8th *International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 2012*. European Language Resources Association (ELRA).

M. Ehrmann, D. Nouvel, and S. Rosset. 2016. Named Entities Resources - Overview and Outlook. In N. Calzolari Conference Chair, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the* 10th *International Conference on Language Resources and Evaluation, Portoro, Slovenia, May 2016.*

Joseph L Fleiss. 1971. 8Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

O. Galibert, L. Quintard, S. Rosset, P. Zweigenbaum, C. Nédellec, S. Aubin, L. Gillard, J-P. Raysz, D. Pois, X. Tannier, L. Deléger, and D. Laurent. 2010. Named and specific entity detection in varied data: The quæro named entity baseline evaluation. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the* 7th *International Conference on Language Resources and Evaluation, Valletta, Malta, May 2010*, pages 3453–3458.

O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard. 2011a. Extended Named Entity Annotation on OCRed Documents : From Corpus Constitution to Evaluation Campaign. pages 3126–3131.

O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard. 2011b. Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In *Proceedings of* 5th *International Joint Conference on Natural Language Processing*, pages 518–526, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

O. Galibert, J. Leixa, G. Adda, K. Choukri, and G. Gravier. 2014. The ETAPE speech processing evaluation. In *Proc. of the* 9th *International Conference on Language Resources and Evaluation (LREC'09)*, Reykjavik, Iceland.

J-N. Gerhard and W. van den Heuvel. 2015. Survey Report on Digitisation in European Cultural Heritage Institutions 2015. Technical report, Europeana/ENUMERATE, June.

G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of the* 6th *International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May – 1 June 2008*, Turkey.

R. Grishman and B. Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proceedings of the* 16th *Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Grover, S. Givon, R. Tobin, and J. Ball. 2008. Named entity recognition for digitised historical texts. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the* 6th *International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May – 1 June 2008.*

S. Van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2):262–279.

T. Huet, J. Biega, and F. Suchanek. 2013. Mining history with Le Monde. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 49–54. ACM.

J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpusa semantically annotated corpus for biotextmining. *Bioinformatics*, 19(suppl 1):i180–i182.

J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. 1999. Performance Measures For Information Extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

C. Neudecker and A. Antonacopoulos. 2016. Making Europes historical newspapers searchable. In *Proceedings of the* 12th *IAPR International Workshop on Document Analysis Systems (DAS2016).*

D. Nouvel, J.-Y. Antoine, and N. Friburger. 2014. Pattern mining for named entity recognition. *LNCS/LNAI Series*, 8387i (post-proceedings LTC 2011).

C. Parada, M. Dredze, and F. Jelinek. 2011. OOV Sensitive Named-Entity Recognition in Speech. In

Proceedings of the 12^{th} *Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 2085–2088, Florence, Italy. International Speech Communication Association (ISCA).

S. Parmesan, U. Scaiella, M. Barbera, and T. Tarasova. 2014. Dandelion: from raw data to dataGEMs for developers. In *Proceedings of the 2014 International Conference on Developers-Volume 1268*, pages 1–6. CEUR-WS. org.

J. Piskorski. 2007. ExPRESS Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of the International Workshop Finite-State Methods and Natural Language Processing 2007 (FSMNLP 2007)*, Potsdam, Germany, September.

D. Rao, P. McNamee, and M. Dredze, 2013. *Multi-source, Multilingual Information Extraction and Summarization*, chapter Entity Linking: Finding Extracted Entities in a Knowledge Base, pages 93–115. Springer Berlin Heidelberg, Berlin, Heidelberg.

A. Ritter, S. Clark, M. Etzioni, and O. Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. J. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 410–414. ÖGAI, September. LThist 2012 workshop.

S. Rosset and C. Grouin. 2011. Entités Nommées Structurées: guide d'annotation QUAERO. Technical report, LIMSI-CNRS.

S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn, and P. Zweigenbaum. 2012. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proceedings of the* 6^{th} *Linguistic Annotation Workshop*, pages 40–48. Association for Computational Linguistics.

C. Sporleder. 2010. Natural language processing for cultural heritage domains. *Language and Linguistics Compass*, 4(9):750–768.

R. Steinberger, B. Pouliquen, and E. van der Goot. 2009. An introduction to the Europe Media Monitor family of applications. In F. Gey, N. Kando, and J. Karlgren, editors, *Information access in a multilingual world Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR)*, Boston, USA, July.

R. Steinberger, B. Pouliquen, M. M. Kabadjov, and E. van der Goot. 2011. JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. In *Proc. of the* 8^{th} *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, Hissar, Bulgaria, September.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the* 13^{th} *Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Vilain, J. Su, and S. Lubar. 2007. Entity Extraction is a Boring Solved Problem: Or is It? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, pages 181–184. Association for Computational Linguistics.