Unveiling Strengths and Weaknesses of NLP Systems Based on a Rich Evaluation Corpus: the Case of NER in French
Alice Millour, Yoann Dupont, Karën Fort, Liam Duignan

HAL Id: hal-04534593
https://hal.science/hal-04534593
Submitted on 5 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Unveiling Strengths and Weaknesses of NLP Systems Based on a Rich Evaluation Corpus: the Case of NER in French

Alice Millour¹, Yoann Dupont², Karën Fort³,4, Liam Duignan¹

¹ LIASD - Pastis, Université Paris 8 Vincennes Saint-Denis (France)
² Lattice (UMR 8094), Université Sorbonne Nouvelle (France)
³ Sorbonne Université (France)
⁴ LORIA, Université de Lorraine (France)

am@up8.edu, yoann.dupont@sorbonne-nouvelle.fr, karen.fort@loria.fr, duignan.liam202@gmail.com

Abstract

Named Entity Recognition (NER) is an applicative task for which annotation schemes vary. To compare the performance of systems which tagsets differ in precision and coverage, it is necessary to assess (i) the comparability of their annotation schemes and (ii) the individual adequacy of the latter to a common annotation scheme. What is more, and given the lack of robustness of some tools towards textual variation, we cannot expect an evaluation led on an homogeneous corpus with low-coverage to provide a reliable prediction of the actual tools performance. To tackle both these limitations in evaluation, we provide a gold corpus for French covering 6 textual genres and annotated with a rich tagset that enables comparison with multiple annotation schemes. We use the flexibility of this gold corpus to provide both: (i) an individual evaluation of four heterogeneous NER systems on their target tagsets, (ii) a comparison of their performance on a common scheme. This rich evaluation framework enables a fair comparison of NER systems across textual genres and annotation schemes.

Keywords: Named Entity Recognition (NER), Evaluation, Textual variation

1. Introduction

Fair evaluation and comparison of NLP models are essential for developers and users. Eval4NLP workshops, initiated in 2020, emphasized this requires extensive work on both the usage of meaningful and interpretable metrics, and the development of quality and unbiased reference resources. Comparing systems is even more complex when multiple annotation schemes coexist. Named Entity Recognition (NER), on which we chose to focus on this work, is a good example of this difficulty. NER is a twofold task that comprises (i) a segmentation step and (ii) a classification (labeling) step, both being determined by the annotation scheme. Because NER is an applicative task, annotation schemes vary between systems, hence producing system-specific outcomes.

To compare the performance of various systems, it is necessary to assess (i) the comparability of their annotation schemes and (ii) the adequacies of the latter to a common reference annotation scheme. The application goals of NER tools are diverse in terms of linguistic contexts and use cases. To provide a reliable evaluation, it is thus necessary to confront their performance with (i) interpretable evaluation metrics and (ii) reference corpora that are unbiased in terms of content.

In this paper, we introduce an evaluation framework based on three dimensions: (i) the precision of the tagset, (ii) the textual genres of the evaluation corpora, and (iii) the evaluation metrics. We use this framework to evaluate four NER systems for French, one rule-based and two neural network-based, both independently and plotted against each other. We show that the gold annotation scheme must be seen as a dimension of the evaluation framework to ensure fair and comprehensive comparison between systems.

The contributions of this paper are the following: (i) a reference corpus of 15 samples of 1 000 tokens split between 6 textual genres categories (prose, poetry, speech, encyclopedic, multisource and informative) and manually annotated with a precise hence flexible tagset of 8 types and 23 sub-types, (ii) mappings between the gold tagset, the target tagsets of the evaluated tools, and the minimal tagset at their intersection (iii) a fair evaluation and comparison of French NER systems.

2. Related Work

Multiple metrics have been developed to evaluate NER. Chinchor and Sundheim (1993) introduced five of them to compare gold and obtained outputs at entity level. Those metrics distinguish strictly and partially correct annotations. During the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003), strict precision, recall and F-score were used. Segura-Bedmar et al. (2013) reused Chinchor and Sundheim (1993) in their evaluation protocol. Tools are evaluated using four types of F-score based on strict or loose matching of types and boundaries. Caubrière et al. (2020) refined the Slot Error Rate variation used in the Quaero shared task (Makhoul et al., 1999; Galibert
et al., 2011) by using hierarchical typing of error, each error type bearing its own weight. Recent works provide frameworks and comparative studies on evaluation reproducibility. Schmitt et al. (2019) evaluate five NER systems for English (StanfordNLP, NLTK, OpenNLP, SpaCy, Gate) on unseen corpora and show that the results are consistently lower than those found in the literature. Palen-Michel et al. (2021) introduce the SeqScore evaluation framework and compare two systems over 10 African languages with a focus on invalid transitions. Yet, these strict comparisons between systems lead to an evaluation on impoverished annotation schemes that match the smallest common tagset (see also Jiang et al. (2016)).

3. Experimental Setup

3.1. Systems

We chose to evaluate four freely available NER systems for French, which exhibit good performance. The first one, CasEN (Friburger and Maurel, 2004), is based on lexical resources, local pattern descriptions and transducers. It was evaluated on the ESLO 1 corpus (Abouda and Baude, 2006) (300 hours of transcribed interviews), and during the ESTER 2 evaluation campaign (Galliano et al., 2009). The second one is the fr_core_news_lg model of SpaCy (Honnil and Montani, 2017), based on neural networks, which gives the best results on the French part of WikiNER (F1-score of 0.84\(^2\)) among off-the-shelf SpaCy models. The third one is the ner-french model of Flair (Akbik et al., 2018), that uses a bidirectional Long short-term memory (LSTM) neural network and contextual embeddings (F1-Score of 0.91 on WikiNER\(^3\)). At last, we use CamemBERT-ner, a camemBERT NER model fine-tuned on the wikitext-fr dataset (F1-score of 0.89\(^4\)). It has been to be noted that both neural network based systems were also trained on the French section of WikiNER.

The coverage and precision of the annotation schemes vary significantly between CasEN (eight types\(^5\)) on the one hand, and SpaCy, Flair and CamemBERT-ner (four types\(^6\)) on the other.

3.2. A New Reference Corpus for French

A fair comparison between multiple tools requires a sufficiently rich gold corpus, both in terms of content and precision of annotations. Since no such
corpus was available for French, we created the first balanced-sample freely available gold corpus for NER in French. 15 samples of 1,000 tokens were selected from 13 sources belonging to six different genres (prose, poetry, speech, encyclopedic, multisource and informative) and available under a free license (see table 1). Their publication period ranges from the 18\(^{th}\) century to the 21\(^{st}\). The main drawback of this classification is the presence of a multisource category, even though individual sentences could have been distributed across other existing categories. We chose for our gold annotation a subset of the Quæro tagset (Grouin et al., 2011) which contains eight types and 23 subtypes. More precisely, only the entities, not the components, were annotated. Annotations were added during two academic years by Master’s students in computational linguistics using WebAnno (de Castilho et al., 2014). The students were given the Quæro annotation guide\(^7\) and were briefly trained before annotating. Each text was annotated by two students, then the annotations were corrected (curated) collaboratively by three NER experts.

We estimated the inter-annotator agreement between the students using Krippendorff’s \(\alpha\) (Krippendorff, 2013). The overall results for the first campaign reached an \(\alpha\) of 0.78±0.06 with 95% confidence interval (0.82±0.03 once the worst annotator has been excluded). The second campaign, in which more difficult genres, such as poetry, were annotated, took more time to correct, with \(\alpha\) of 0.35±0.43\(^8\).

Some of the hardest entities include rare elements such as historical events uneasy to identify (eg. “conspiration des poudres”, - the “Gunpowder Plot”) or religious items (eg. “Pentateuque” - “Pentateuch”, “les Écritures” - “Scriptures”). What is more, we distinguish two types of ambiguity: - apparent ambiguity which can be resolved with context. These include for instance a series of mythological names which could either refer to LOC or PERS (eg. “Lesbos” or “Venus”). These probably wrongly annotated by some of the students in reason of the texts’ complexity coupled with attention slips. - intrinsic ambiguity which could either be resolved by a very careful reading of the annotation guidelines, or required a decision. This is the case for entities such as “tourism office”, which can be interpreted as LOC or ORG. We documented our choices along curation to ensure consistency. This documentation is made available with the corpus.

\(^3\)See: https://huggingface.co/flair/ner-french.
\(^4\)See https://huggingface.co/Jean-Baptiste/camembert-ner.
\(^5\)amount, prod, org, loc, time, func, per, event.
\(^6\)ORG, LOC, PER, MISC.
\(^8\)These agreements were computed using the Rmisc library, see: https://www.rdocumentation.org/packages/Rmisc/versions/1.5.
After curation, the gold corpus contains 1,124 entities annotated with 31 tags. The distribution across the main types is given in Table 2.

### 4. Comparing Apples and Oranges

The three annotation schemes used in this work (gold, CasEN, and WikiNER) are not directly comparable. As a consequence, we measured the adequacy of the schemes and built intermediate mappings between the three tagsets.

#### 4.1. Matching Annotation Schemes

**Meta-tags and Evaluation Corpus** To evaluate each tool individually, we created two series of meta-tags at the intersection between (i) the gold scheme (eight types that cover 1,124 NEs) and the tagset used by CasEN (eight types that cover 1,088 reference NEs), and (ii) the gold scheme and the minimal annotation scheme, which was used in the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) and is used by the three other systems (four types that cover 819 of the gold NEs).

To enable a basic comparison between tools, we also created a lookup table between the CasEN annotation scheme and this minimal one. We call the first intersecting tagset (i) Meta-CasEN, and the second one (ii) Meta-WikiNER. The corresponding mappings are provided in Appendices A.1 and A.2. Entity counts per tagset are given in table 3.

The Quæro annotation scheme being the richest tagset used in this study, the transposition of the gold corpus to the two meta-tagsets causes a diminution of the number of NEs. For example, the entities amount and date are excluded from the minimal evaluation corpus.

#### Comparing Annotation Schemes

Since we evaluate systems on an annotation scheme that differs from the one they were trained with and for, it is necessary to assess their adequacy, both syntactically and semantically. To do so, we used a WikiNER sample on which we compared the semi-automatically created WikiNER gold standard (used to train both SpaCy and Flair, and to finetune CamemBERT-ner) with the two variants (Quæro based and minimal) of our gold standard.

Table 4 shows the high precision between the tagsets: a WikiNER NE is generally a Quæro NE. Recall increases naturally as the tagset gets poorer. The main difference we observe between annotation schemes comes from inconsistent NE segmentations due to the semi-automatic annotation of the WikiNER corpus. Such inconsistencies include for example the variable inclusion of the determiner before the NE, or the NE being either split into two annotations or considered as single one. The latter is especially true for locations such as “Chine du Nord”.

Those inconsistencies generate boundary and labeling errors on NEs. There is a high consistency between the minimal and the WikiNER annotation schemes (dates and amount are not labeled ; prod annotations become MISC). Still, part of the mismatching annotations are due to apparent inconsistencies between the two annotation schemes. For instance, music bands are annotated ORG in WikiNER and pers.coll in Quæro, “New York Times” is annotated ORG and prod.media. In fact, these differences between the annotation schemes are a direct consequence of the limitations of a coarse-grained 4 label tagset. Some equivalences are up to debate. For instance, we chose to match func with MISC, but the CoNLL-2003 guide (along with an examination of the WikiNER annotation) does not allow us to compare the correspondence between NE spans, which appears to be high.

Table 1: Description of the manually annotated corpus.

<table>
<thead>
<tr>
<th>Source (identifier)</th>
<th>Period</th>
<th>Genre</th>
<th>Nb. sentences (Nb. tokens)</th>
<th>Licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traité sur la tolérance, Voltaire (42131-0)</td>
<td>18th</td>
<td>prose</td>
<td>40 (1,020)</td>
<td>Project Gutenberg</td>
</tr>
<tr>
<td>Le Ventre de Paris, Émile Zola (pg6470)</td>
<td>19th</td>
<td>prose</td>
<td>51 (1,002)</td>
<td>Project Gutenberg</td>
</tr>
<tr>
<td>L’Homme qui plantait des arbres, Jean Giono (Wikisource)</td>
<td>20th</td>
<td>prose</td>
<td>53 (1,013)</td>
<td>Project Gutenberg</td>
</tr>
<tr>
<td>Les Fleurs du Mal, Baudelaire (pg6099-96)</td>
<td>19th</td>
<td>poetry</td>
<td>30 (1,014)</td>
<td>Project Gutenberg</td>
</tr>
<tr>
<td>Œuvres d’Arthur Rimbaud - Vers et proses (56708-0)</td>
<td>19th</td>
<td>poetry</td>
<td>52 (1,027)</td>
<td>Project Gutenberg</td>
</tr>
<tr>
<td>UD French GSD</td>
<td>21st</td>
<td>multirources</td>
<td>35 (1,021)</td>
<td>CC BY-SA 4.0</td>
</tr>
<tr>
<td>Sequoia (Candito and Seddah, 2012)</td>
<td>21st</td>
<td>multirources</td>
<td>44 (1,002)</td>
<td>CC BY-SA 4.0</td>
</tr>
<tr>
<td>French Question Bank (Seddah and Candito, 2016)</td>
<td>21st</td>
<td>multirources</td>
<td>102 (1,006)</td>
<td>CC BY-SA 4.0</td>
</tr>
<tr>
<td>APIL (office du tourisme Othe-Armance)</td>
<td>21st</td>
<td>information</td>
<td>29 (1,002)</td>
<td>CC BY-SA 4.0</td>
</tr>
<tr>
<td>Wikinews</td>
<td>21st</td>
<td>encyclopedia</td>
<td>46 (1,024)</td>
<td>CC BY-SA 2.0</td>
</tr>
<tr>
<td>L’Est Républicain (ATILF and CLLE, 2020)</td>
<td>21st</td>
<td>information</td>
<td>40 (1,000)</td>
<td>CC BY-SA 2.0</td>
</tr>
<tr>
<td>French WikiNER</td>
<td>21st</td>
<td>spoken</td>
<td>213 (3,081) (3 samples)</td>
<td>CC BY-SA 4.0</td>
</tr>
<tr>
<td>Hhahpsodie (Lacher et al., 2014)</td>
<td>21st</td>
<td>spoken</td>
<td>213 (3,081) (3 samples)</td>
<td>CC BY-SA 4.0</td>
</tr>
</tbody>
</table>

Table 2: Type distribution across the corpus.

<table>
<thead>
<tr>
<th>amount</th>
<th>event</th>
<th>func</th>
<th>loc</th>
<th>org</th>
<th>pers</th>
<th>prod</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>9%</td>
<td>3%</td>
<td>8%</td>
<td>31%</td>
<td>8%</td>
<td>20%</td>
<td>6%</td>
<td>15%</td>
</tr>
</tbody>
</table>

9Differences observed in LOC and PROD counts result from filtering out loc.add.elec and prod.doctr entities in the mapped corpora (see Appendix A.1.)

10“Northern China”.

11The ambiguity with org.ent tag is solved in Quæro, section 1.3.4. (Rosset et al., 2011).
Table 3: Entity counts per tagset.

<table>
<thead>
<tr>
<th>Tagset</th>
<th>strict</th>
<th>partial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quæro (n/t)</td>
<td>P 0.91</td>
<td>R 0.56</td>
</tr>
<tr>
<td></td>
<td>F 0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P 0.95</td>
<td>R 0.58</td>
</tr>
<tr>
<td></td>
<td>F 0.72</td>
<td></td>
</tr>
<tr>
<td>Quæro</td>
<td>P 0.83</td>
<td>R 0.51</td>
</tr>
<tr>
<td></td>
<td>F 0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P 0.88</td>
<td>R 0.53</td>
</tr>
<tr>
<td></td>
<td>F 0.66</td>
<td></td>
</tr>
<tr>
<td>minimal (n/t)</td>
<td>P 0.91</td>
<td>R 0.83</td>
</tr>
<tr>
<td></td>
<td>F 0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P 0.95</td>
<td>R 0.86</td>
</tr>
<tr>
<td></td>
<td>F 0.90</td>
<td></td>
</tr>
<tr>
<td>minimal</td>
<td>P 0.84</td>
<td>R 0.79</td>
</tr>
<tr>
<td></td>
<td>F 0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P 0.89</td>
<td>R 0.84</td>
</tr>
<tr>
<td></td>
<td>F 0.87</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Comparison between WikiNER and our gold annotations. "n/t" (for no types) means that the types were ignored for the comparison.

lead to a clear-cut conclusion. The case of multiple annotations LOC+ORG, specific to metonimic NEs, leads to a partial match and shows the limitations of the WikiNER tagset.

The CasEN tagset is closer to the Quæro tagset, both being derived from the ESTER 2 corpus tagset. The only difference we noticed is that CasEN uses the type ref for both numerical references (for example "[1]") and for book titles, while only the latter is annotated in Quæro.

4.2. Results

We compared the results of the systems according to different evaluation setups illustrated on Figure 1. The graph on the left shows the strict F-scores reached by genre and on the whole corpus, while the graph on the right gives the absolute number of correct annotations and error types (namely substitution, insertion and deletion, as defined by Galibert et al. (2011)) on the whole corpus only. Because the reference annotation scheme is rich, we can provide two series of results for CasEN. The first one (labeled as CasEN) is an autonomous evaluation of the system, evaluated on its target tagset, the second is a reduction of its output to the minimal WikiNER scheme (Minimal-CasEN) Globally (with all genres mixed), all systems show similar F-scores, with CamemBERT-ner outperforming the three other systems.

We can see the value of using multiple metrics instead of a unique one. Single metric benchmarks do not convey how systems behave on actual data: for instance the CasEN system is prone to insertion, deletion and boundary errors but presents the highest score for correct annotation and shows the best performance for the labeling subtask. What is more, the left part shows that CasEN outperforms Flair and SpaCy on literary genres. These systems indeed tend to fluctuate with the textual genres, with better performances on the "Multisource" and "Encyclopedic" axes, while the performance of CasEN is stable.

The figure also shows that the systems have comparable error profiles. This suggests a potential subset of entities that are harder to properly annotate.

5. Conclusion

We detailed in this paper the work that is required to compare tools developed for the same task. We showed the benefits of building a rich evaluation framework based on an adequate gold reference and explicit metrics. The reference corpus and the mappings between annotation schemes that this comparison requires are freely available. The different textual categories of the gold corpus allows a more accurate evaluation of off-the-shelf systems on various textual genres. This gives an insight on the robustness of these systems, as some genres were absent from their training data.

This study also confirms that challenging systems on a variety of corpora reveals the over-fitting tendencies of some of them.

6. Limitations

We intend to complete the evaluation using the methodology proposed by Fu et al. (2020), who, in order to get a better understanding of errors and possible improvements, use new metrics using NE attributes such as length in number of words, or token frequency and ambiguity. Building interpretable metrics based on meaningful NE traits would allow for a better understanding of the actual performance of the systems. Our experiment could be enriched through a finer per-type analysis.

At the time of this article, our gold corpus contains 15,200 tokens, which is about 40 % of the test set of the reference corpus for written French, the French Treebank (Abeillé et al., 2003), as split in Crabbé and Candito (2008). As for sampling, 12See section 1.4 of the Quæro annotation guide (Rosset et al., 2011).

13See https://github.com/aliciemillour/FENEC.
while each single sample is of the same size, their
distribution across genres is not as well balanced,
poetry and encyclopedia being under-represented.
A perspective of our work is therefore to extend the
corpus and improve its balance across genres at
the same time. This work has already started, with
the help of our Masters’ students.

7. Acknowledgments
We thank all the annotators for their work.

Bibliography


Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard.


A. Appendix

A.1. Minimal meta-tagset used to evaluate SpaCy and Flair

WikiNER tagset

ORG

LOC

PER

MISC

Quæro tagset

org.adm

org.ent

loc.phys.geo

loc.adm.reg

loc.phys.hydro

loc.oro

loc.fac

loc.add.elec*

loc.other

loc.adm.nat

loc.adm.town

loc.adm.sup

loc.phys.astro

pers.ind

pers.coll

prod.object

prod.art

prod.rule

prod.media

prod.software

prod.award

func.ind

func.coll

event

Meta-tagset

ORG

LOC

PER

MISC

*loc.add.elec are filtered out from the LOC annotations since they are not annotated as such by the systems

** This type groups any annotation that could be annotated as MISC
A.2. **Minimal meta-tagset used to evaluate CasEN**

* the nationality tag was excluded

** loc.add.elec are filtered out from the LOC annotations since they are not annotated as such by the systems