WikiNER-fr-gold: A Gold-Standard NER Corpus

Danrun Cao, Nicolas Béchet, Pierre-François Marteau

WikiNER-fr-gold: A Gold-Standard NER Corpus

Danrun Cao[1,2], **Nicolas Béchet**[1], **Pierre-François Marteau**[1]
(1) Univ. Bretagne Sud, CNRS, IRISA
Rue Yves Mainguy, 56000 Vannes, France
{danrun.cao, nicolas.bechet, pierre-francois.marteau}@univ-ubs.fr
(2) OctopusMind
2 Pl. Saint-Pierre, 44000 Nantes, France
d.cao@octopusmind.info

Abstract

We address in this article the the quality of the WikiNER corpus, a multilingual Named Entity Recognition corpus, and provide a consolidated version of it. The annotation of WikiNER was produced in a semi-supervised manner i.e. no manual verification has been carried out *a posteriori*. Such corpus is called silver-standard. In this paper we propose WikiNER-fr-gold which is a revised version of the French proportion of WikiNER. Our corpus consists of randomly sampled 20% of the original French sub-corpus (26,818 sentences with 700k tokens). We start by summarizing the entity types included in each category in order to define an annotation guideline, and then we proceed to revise the corpus. Finally we present an analysis of errors and inconsistency observed in the WikiNER-fr corpus, and we discuss potential future work directions.

Keywords: Annotated corpus, resource production, Named Entity Recognition (NER), French

1. Introduction

In Natural Language Processing (NLP), Named Entity Recognition (NER) is a task that focuses on identifying entities within unstructured text. The goal of an NER system is to locate nominal phrases referring to an entity and assign them a category from a predefined list. This phrase is referred to as the mention of an entity, and defined as a series of one or multiple consecutive tokens corresponding to one specific and unique entity. A token is defined as a continuous sequence of non-empty characters, representing the minimal unit during the automatic processing of textual data. The NER task has a dual objective: determining the boundaries of a mention and categorizing the entity that is mentioned.

Training an NER system requires an annotated corpus. For French language there exists annotated corpora, but few are freely available. The French Treebank (Abeillé et al., 2003), composed of articles from the newspaper *Le Monde* (1990-1993), is a corpus for French syntactic analysis. It served as the basis for one of the first corpora dedicated to French NER, as presented in (Sagot et al., 2012). This corpus consists of 5,890 sentences with a total of 11,636 entities. Another usable corpus is Europeana Newspapers (Neudecker, 2016), which contains digitized newspaper articles processed with OCR tools. It is a multilingual corpus, and the French part contains 12,551 sentences. However, this corpus requires significant correction work before use, because many OCR-related errors remain disseminated in the corpus. The FENEC corpus (Millour et al., 2022) was created from six text genres (prose, poetry, journalistic text, encyclopedia, speech, and multi-sources). This corpus contains 11,149 tokens and 875 entities and was annotated following the Quaero schema (Rosset et al., 2011).

The largest NER corpus we have identified is WikiNER (Nothman et al., 2013), an encyclopedic corpus covering ten languages, including French. Several open-source NER tools have been trained on this corpus, such as spaCy, Flair (Akbik et al., 2019), and Spark NLP. This corpus consists of sentences extracted from Wikipedia articles, annotated with named entities. It covers four types of entities: person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). All ten sub-corpora have the same size, comprising approximately 3.5 million tokens, making it a very substantial dataset. The annotations were produced in a semi-supervised manner, and there was no manual verification for the corpus. Therefore, it is considered a silver standard corpus.

In this article, we describe the manual correction process implemented to create a gold standard version of WikiNER. We will refer to this new corpus as WikiNER-fr-gold in the following discussion. This work involved manual correction of 20% of the French portion of WikiNER, which we will refer to as WikiNER-fr. WikiNER-fr-gold comprises 26,818 sentences and approximately 700,000 tokens. These data were randomly selected from the original corpus.

The paper is organized as follows. In section 2, we will provide a brief overview of the production of WikiNER annotations to highlight the origin of typical errors. Section 3 will present the observed

errors along with the correction choices we have made. Finally, in section 4, we will present the future works.

2. Production of the original annotations of WikiNER

The original annotations of WikiNER were produced using hyperlinks of Wikipedia articles. If there exists a Wikipedia page corresponding to an entity mentioned in a sentence, then the phrase describing that object would be linked to its Wikipedia page via a hyperlink. This linkage can be exploited in a reverse way: the text of a hyperlink helps identify an object, which matches the definition of a named entity. The boundaries of the hyperlink naturally serve as those of the mention. It remains simply to project the category of the object onto the mention. Annotation of the original corpus was thus carried out in two steps: the classification of Wikipedia pages, and the annotation of mentions within Wikipedia articles.

Firstly, for each of the 10 languages, the authors created a training corpus to train a classification model. The French corpus consists of approximately 2,500 articles. The annotations follow an extended version of the annotation schema of the BBN corpus (Brunstein, 2002). Next, the authors compared three classification strategies to find the best solution. A classifier was trained using each of these strategies, and the authors reported precision, recall, and F1 score using 10-fold cross-validation. The best method was Logistic Regression with an average F1 score of 94%. It was then used for the classification of the remaining Wikipedia pages.

Then, the categories of Wikipedia pages were projected onto their hyperlinks occurring in other Wikipedia pages in order to make the initial annotations. In Wikipedia, only the first occurrence of an entity receives a hyperlink. The authors proposed several inference strategies in order to retrieve the other mentions in the remaining text. First a list of potential mentions was created for each entity. This list is generated from hyperlinks and redirections to corresponding Wikipedia pages. Not every element of this list is eligible as a mentioned candidate, obviously. The authors then proposed several criteria to filter non-conforming elements. Four rigor levels are defined by varying criteria combinations. A higher level of rigor corresponds to a more strict filtration. Annotation quality may be higher but at the cost of a reduced variety level of mentions. In total, five variants of the corpus are proposed, each with 3.5 million tokens.

In our study, we chose WIKI-2, the version produced with level-2 filtration. It represents a good compromise between annotation quality and entity coverage. Table 1 displays the token count for each entity type in the corpus. Each token can only receive a single label as its entity type and can belong to only one entity.

Entity	PER	LOC	ORG	MISC
Token count	129,978	155,565	45,443	81,594

Table 1: Token count of each entity type in WIKI-2

3. Corpus review

3.1. Entity category definition

The annotation scheme serves to clarify how the categories were defined. We propose summarizing this annotation schema by presenting the types of entities included in each category. Table 3.1 provides a comprehensive list of entity types by category, with reference to some examples.

3.2. Annotation format and tool

The annotations are formatted in BIOES format. Within each entity, we distinguish the beginning (B), inside (I), and end (E) of the entity. This format helps highlight the boundaries of entities. For example, in "*général de Gaulle (General de Gaulle)*", the three tokens receive the labels B-PER, I-PER, and E-PER, respectively. For entities consisting of a single word, we use the label S (for single). So, the entity "France" is annotated as S-LOC. Tokens outside entities are labeled as O, indicating they are not part of any entity. There are a total of 17 formatted labels.

We use the labeling tool provided by (anonymous reference). The advantage of this tool lies in its ability to customize candidate labels and their visual representation. Thanks to this, a color scheme could be defined that facilitates the understanding of the category and boundaries of the entity. Figure 1 provides an overview of the tool's interface.

3.3. Error analysis and correction

During the corpus review, we observed very few clear-cut errors, meaning mentions that do not correspond to either an entity or a Wikipedia page. Most errors are recurring, and we can easily trace their origins in the annotation generation process. In the following paragraphs, we present these errors grouped by their nature. We then explain the corrections made accompanied by examples.

We would like to insist on the fact that the objective of our work is to solely standardize annotations and correct errors. We do not question the logic of the original annotation choices. Thus, as a principle, we do not change the category assigned to an entity unless it is an indisputable error (for example, annotating "*France*" as MISC). In cases of incoherence, when an entity receives multiple categories, we refer to the annotation of other entities of the same

Category	Entity type	Example
LOC	Country and region	*France*, *Loire Atlantique*
	Iconic building	*Gare Montparnasse*, *Tour Eiffel*
	Natural landscape	*Seine*, *Alpes*
	Transport lines and networks	*TGV Est*, *RER A*
	Celestial bodies	*Soleil (Sun)*, *Alpha Centauri*
PER	Name and family	*Staline*, *Maison d'Orange*
	Fictional characters	*Zeus*, *Indiana Jones*
	Nationality and ethnicity	*(les) Français (the French)*, *(les) Aztèques (the Aztecs)*
ORG	Organization, institution	*ONU (United Nations)*, *Fonds monétaire international (International Monetary Fund)*
	Government bodies	*Assemblée Générale (General Assembly)*, *Parlement Irlandais (Irish Parliament)*
	Political parties	*UMP*, *Parti communiste chinois (Chinese Communist party)*
	Companies	*Microsoft*, *EDF (Electricity of France)*
	Sports teams	*Bulls de Chicago (Chicago Bulls)*, *(équipe) France (French national team)*
	Musical bands	*les Beatles*, *AC/DC*
	Higher education institutions	*Université de Paris*, *Université de Californie à Berkeley (UC Berkeley)*
	Military organizations	*Armée Rouge (Red Army)*, *US Marine Corps*
MISC	Titles of works	*La Joconde (Mona Lisa)*, *Bible*
	Events	*Seconde Guerre Mondiale (World War II)*, *Jeux Olympiques (Olympic Games)*
	Historical periods and regimes	*Dynastie Qing (Qing Dynasty)*, *Grèce antique (Ancient Greece)*
	Software and hardware	*(langage) Python*, *PS5*
	Conventions and documents	*Édit de Nantes (Edict of Nantes)*, *(la) Constitution*
	Ships and rockets	*HMS Triumph*, *Ariane 2*
	Brands	*Land Rover*, *TGV*

Table 2: Entity types by category with examples

type and to their corresponding Wikipedia articles to decide whether or not a modification should be made. Also, it is important to note that this review is only applied to entities that have already been identified. We do not add entities unless there is an obvious omission, such as a country name that wasn't annotated.

3.3.1. Inconsistent definition of hyperlinks

The hyperlinks in Wikipedia are manually created by many contributors. There may be a lack of agreement on hyperlink standards, which can result in the generation of inconsistent annotations. For example, in the phrase "*la France (France)*", some link the word "*France*" to the corresponding page, while others also include the article "*la*" in the mention. As a result, in the corpus, both the mentions "*France*" and "*la France*" exist for the same entity "*France*". Similarly, appositions can lead to inconsistencies. The mention "*ville de Lyon (city of Lyon)*" was seen associated with the entity "*Lyon*" instead of the token "*Lyon*" by itself.

Aside from redundant mentions, there exist also incomplete mentions. For example, in the mention "*Coupe du monde (World Cup)*", sometimes only the word "*Coupe (Cup)*" is annotated. This phenomenon is especially common with nested entities (entities that contain other entities). Take the example of "*comté de Mortain (County of Mortain)*", a medieval county centered around the town of Mortain. The entire mention should receive the LOC label, but instead only the town "*Mortain*" has been annotated.

For this type of error, it will suffice to simply remove redundant parts and add missing ones. As a general rule, articles, appositions, and descriptions are removed from the mention, except in two cases. The first case is when they are part of the entity's

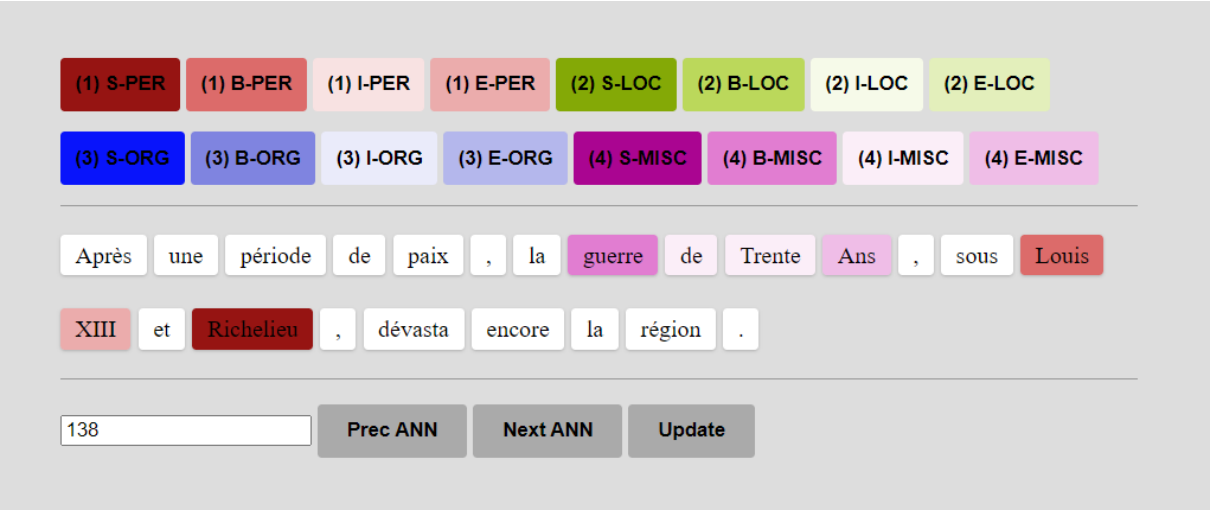

Figure 1: Overview of the labeling tool

name or its conventional appellation. For example, "*Le Havre*" remains "*Le Havre*". "*Général de Gaulle (General de Gaulle)*" also remains a complete mention, although "*général (general)*" is not part of the name. On the contrary, "*le roi Louis XIV (King Louis XIV)*" becomes "*Louis XIV*" because it is understood that he is a king without specifying it. The second case is when their presence is essential to avoid ambiguity. Consider the pair "*ville de Bruxelles (city of Brussels)*" and "*Région de Bruxelles (Region of Brussels)*". The first refers to the city of Brussels, while the latter refers to a region in Belgium of which Brussels is the capital. This also applies in the PER (Person) category, such as "*de Saint-André*," which refers to the journalist Alix de Saint-André, and "*maréchal de Saint-André (Marshal de Saint-André)*", corresponding to the marshal Jacques d'Albon de Saint-André.

It could be difficult to differentiate a nested entity from a false annotation comprising the entity's description. In cases of doubt, we refer to Wikipedia for clarification. We first check the entity with the widest scope to see if there is a Wikipedia page with a title corresponding to the complete mention. If such a page exists, we retain that entity and annotate it. If not, we reduce the mention to the smaller entity and repeat the verification process. There are instances where nested entities have only certain components annotated. In these cases, we try to complete them as much as possible.

3.3.2. Hyperlinks non conforming to the definition of a named entity

Two criteria must be fulfilled for a phrase to be recognized as a mention of a named entity: (1) it must be of **nominal** nature and (2) it must refer to a **specific** and **unique** real-world object. However, Wikipedia pages and hyperlinks do not have such rules. For example, there exists a page "*Relations entre la Chine et le Tibet durant la dynastie Ming (Relations between China and Tibet during the Ming Dynasty)*". But it is not considered an entity since the relationship between two regions is not a clear and precise concept. A hyperlink leading to this page was placed on the phrase "*La Dynastie Ming patronnait l'activité religieuse du Tibet (The Ming Dynasty sponsored religious activity in Tibet)*". The page was annotated as MISC, hence the phrase inherited the same annotation. Aside from the false annotation of the Wikipedia page, a complete phrase cannot be considered a named entity. So we remove the annotation on the phrase, and annotate only the entities "*Dynastie Ming (Ming Dynasty)*" and "*Tibet*". Similarly, "*histoire de la Chine (history of China)*" and "*liste de communes de France (list of municipalities in France)*" are not considered named entities.

Another peculiarity of named entities is that their interpretation depends on context. For example, in a general context, "*Cité Interdite (Forbidden City)*" refers to the ancient royal palace in China. Thus it is annotated as LOC. However, there is also a page about a film bearing the same name. Mentions related to this sense should be annotated as MISC. Furthermore, some entities cannot be interpreted without context. In the sentence "*Sa mère meurt d'un cancer de l'estomac le 15 septembre 1821 (Her mother died of stomach cancer on September 15, 1821)*" from the page "*Charlotte Brontë*", "*Sa mère (Her mother)*" receives a hyperlink to the page "*Maria Brontë*". "*Sa mère*" is therefore annotated as PER. However, this inference is valid only within the original context, i.e., in the article presenting Charlotte Brontë. Without this context, "*sa mère*"

cannot be associated with a specific person. Therefore, this phrase is not considered a named entity, and its annotation is removed.

3.3.3. Entities of complex nature

Certain entities can be challenging to categorize due to their complex nature, especially geopolitical entities. For example, in Wikipedia, the "*Empire Britannique (British Empire)*" is defined as "*l'ensemble des territoires qui, sous des statuts divers [...] ont été gouvernés ou administrés du XVI au XX siècle par l'Angleterre, puis le Royaume-Uni*" (the set of territories that, under various statuses [...] were governed or administered from the 16th to the 20th century by England, then the United Kingdom). If we consider it as a group of colonies, then the entity can be seen as a geographical concept and annotated as LOC. However, there is also an organizational and hierarchical structure between the United Kingdom and its colonies. In this sense, it is also appropriate to annotate it as ORG. This discussion can apply to other entities of the same kind, such as "*Empire romain (Roman Empire)*", "*Grèce antique (Ancient Greece)*", and "*Allemagne nazie (Nazi Germany)*".

Now consider "*Carthaginois (Carthaginians)*" in the sentence "*Les Carthaginois prennent d'abord la ville de Messine*" (The Carthaginians first take the city of Messina). Annotating it as PER seems right since it refers to the people of the Carthaginian civilization that occupied Messina. However, the entire population did not participate in the war, but rather the Carthaginian army. Following this logic, "*Carthaginois*" should be annotated as ORG. But once again, army or people, war is an act that involves two nations. Therefore, it would also be possible to annotate this entity as LOC.

For such entities, it is difficult to assign a single label, and the annotation choices of the authors can be easily contested. We try to follow the original annotation schema when dealing with them. If the entity appears elsewhere in the corpus, we adopt the same label. If not, we refer to other entities of the same type and assign a label that we find most appropriate. One special case regards nationalities or ethnicities, such as "*Carthaginois*". The annotation of these entities is highly diverse, all four labels can be found. We have made the decision to annotate them all as PER, but the debate remains open.

4. Conclusion

We have presented WikiNER-fr-gold, a gold-standard NER corpus in French. The corpus consists of 20% of the WikiNER-fr corpus, randomly sampled, which is then subjected to manual revision. Our goal was to standardize and homogenize the annotations while following the original annotation schema as much as possible.

One limitation in this work is the lack of comparison with other annotation schemes. For example, titles such as "*Duc de Bretagne (Duke of Brittany)*" are considered an entity only when it refers to one specific person deductible from the sentential context. This choice was made in coherence with the definition of a named entity. However in the Quaero corpus, they are annotated PER, since a title is associated with a person, even when we do not know precisely which one. It would have been interesting to compare the handling of such cases in other corpora, and if possible, to hear their authors' explanation on annotation choices.

In future works, we will perform a more comprehensive assessment of WikiNER's annotations regarding other NER corpora, with the goal of a revision of entity categorization. This could be the occasion, for example, to revisit the annotation of geopolitical entities. Ideally, this corrective process would be applied to the entire corpus. Some of the corrections can be automated, especially for certain recurring errors. Redundant articles, for instance, can be easily identified using rules and lexicons. We can also solicit the Wikipedia API to facilitate the detection of embedded entities. Furthermore, it would be interesting to implement an active learning system during the correction. We can train an assistant model that takes into account previously encountered errors and then identifies potential erroneous mentions. The new annotation guidelines will be distributed with the corpus to keep the task of expanding WikiNER-fr-gold open and active. Finally, we will extend the revision work to the entire WikiNER-fr, and eventually to other languages.

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20, pages 165–187. Springer Netherlands, Dordrecht.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *NAACL 2019*, pages 54–59.

Ada Brunstein. 2002. ANNOTATION GUIDELINES FOR ANSWER TYPES.

Alice Millour, Yoann Dupont, Alexane Jouglar, and Karën Fort. 2022. FENEC: un corpus à échantillons équilibrés pour l'évaluation des entités nommées en français. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 82–94, Avignon, France. ATALA.

Clemens Neudecker. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. Entités nommées structurées : guide d'annotation Quaero.

Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, volume 2, pages 535–542, Grenoble, France. ATALA/AFCP.