UFEL: a By-design Understandable and Frugal Entity Linking System for French Microposts

Vivien Leonard, Jean-Yves Antoine, Béatrice Markhoff

UFEL: a By-design Understandable and Frugal Entity Linking System for French Microposts

Vivien Leonard[1,2][0009−0006−2274−0394], Béatrice Markhoff[3][0000−0002−5171−8499], and Jean-Yves Antoine[2,4][0000−0002−6028−1663]

[1] ATOS France
[2] LIFAT, University of Tours, France
[3] UMR 7324 CITERES, CNRS University of Tours, France
[4] LIFO, University of Orleans, France
{beatrice.markhoff,jean-yves.antoine}@univ-tours.fr
vivien.leonard@etu.univ-tours.fr

Abstract. Entity Linking (EL) is a critical task in Information Extraction (IE) that involves associating Named Entities (NEs) mentioned in text with their corresponding entity in a Knowledge Base (KB). This intersection of Natural Language Processing (NLP) and Semantic Web Exploitation (SWE) has mainly been investigated on large texts, such as the newswire. However, the shift towards analysing microposts - short, informal social media content — has revealed significant shortcomings in the performance of conventional EL methods. We introduce UFEL, a novel zero-shot strategy for EL task in the context of microposts conversations. Our methodology capitalises on open Semantic Web resources, including the Wikipedia and Wikidata APIs, DBPedia and Wikidata as KBs, and operates under the assumption that consecutive NEs within a micropost exhibit a high degree of semantic interconnection. Furthermore, the simplicity of our scoring mechanism makes our solution efficient and easily understandable while reaching 75% F1-score, setting a new state of the art in the context of French microposts data. We evaluate UFEL with respect to three other systems, including ReFinED, a deep learning system from Amazon.

Keywords: Microposts · Entity Linking · Open Knowledge Graphs.

1 Introduction

In the contemporary digital landscape, micropost platforms (e.g. X, Mastodon, etc.) have emerged as large sources of textual data, presenting opportunities for knowledge extraction and relationship analysis. Every day, millions of users contribute to this ever-expanding universe of microposts, succinct messages that, despite their brevity, are rich in information and societal insights. This is especially relevant in our case because this work will be part of a software suite developed by ATOS, dedicated to social network analysis, and used in various company projects. One of those projects, *Cloud Platform For Smart City*

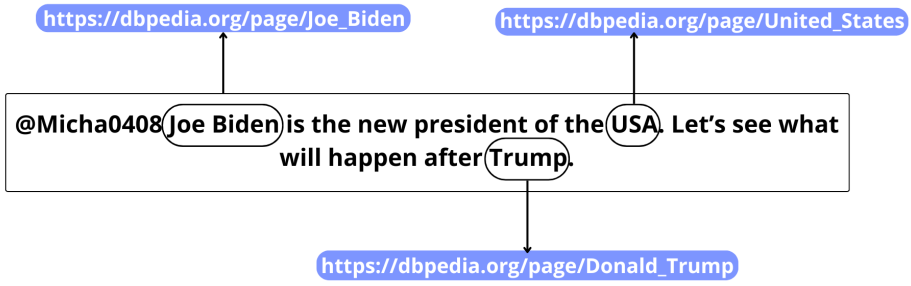

Fig. 1. Example of the EL task including the NER and NEL subtasks

(CP4SC)[5] aims at aggregating data about cities, including microposts that can offer real-time insights into urban dynamics, public opinion, or trends. The volume and dynamic nature of the data generated on micropost platforms make it an important type of data source for researchers, analysts, and urban planners seeking to gather and analyse public sentiment, monitor urban trends, and respond to real-time events in the cityscape. It is important for ATOS, and particularly for the CP4SC project, to be able to deal with microposts in French. As no freely available microposts corpus exists for French, we built one for performing our experiments.

The very characteristics that make micropost data so valuable also render its analysis challenging. The brevity of microposts, often accompanied by idiosyncratic language, abbreviations, and a lack of context, poses significant challenges for traditional text processing techniques. In addition, a rapid and continuous stream of data requires efficient and scalable processing methods to capture and analyse information in a timely manner.

One of the most important tasks when it comes to analysing vast quantities of texts is the task of Entity Linking (EL). The objective of this task is to link mentions from a text to the corresponding entry inside a Knowledge Base (KB). In the rest of this article, we use the term *mention* to refer to Named Entities (NEs) in texts, and we use the term *entity* to refer to an entity inside a KB, i.e., a URI identifying an entity. EL involves two key steps. First, the extraction of mentions within the considered text, a task known as Named Entity Recognition (NER). Second, the identified mentions are linked to their corresponding entities in a KB, a process called Named Entity Linking (NEL). NEL can further be divided into two subtasks, i.e. Candidate Generation (CG), the generation of a set of candidate entities for the considered mention, and Candidate Selection (CS), the decision process that will select from the set of candidates the relevant entity. An example of result for the entire process is given in Figure 1.

[5] `https://eviden.com/industries/public-sector-and-defense/cloud-platform-for-smart-cities/`

The performance of EL algorithms tends to deteriorate significantly when applied to micropost data [6], a decline that can be traced back to the distinct nature of the data being processed. Recently, numerous Deep Learning (DL) approaches have been developed to address this task [26] or to accommodate the specificities of X data [2]. However, these approaches come with their set of limitations, including the need for very large annotated resources to learn from, complexity, and lack of understandability.

Also from an ecological and social perspective, DL approaches are limited. They are based on the assumption that there is an infinite amount of resources on Earth (materials, electricity, water) and that the amount of energy necessary to use and develop this type of method will always be available. For years, this assumption is not true any more, and lighter approaches, using open semantic digital commons and particularity in the data and the the context in which they are produced, must be at the heart of the development of new approaches. This becomes even more true these days, due to the development of Large Language Models (LLMs), requiring more and more resources [3,16,17,4,29].

In response to the challenges exposed by microposts and the lack of large annotated resources, especially for French, we have chosen to build a method that is both streamlined and easily interpretable, yet still maintaining robust performance, based on the standard EL pipeline. The mention recognition step relies on an existing basic pre-trained DL resource because of its robustness against irregular language uses in microposts. Our contribution focusses on the NEL task and consists of a CG step based on the Wikipedia API [6] and the Wikidata API[7], and a CS step based on a scoring formula combining features. Thanks to the high multilingualism provided by these semantic Web resources which we exploit in an important use case, our proposal is not limited to French, and can be adapted to other natural languages (this is a future work for us to do it).

This proposal is one brick of a broader framework whose aim is to build a knowledge graph mirroring a microposts conversation, as illustrated in Figure 2. Methods for querying, discovering, and analysing knowledge graphs could then be leveraged to observe the language and functioning of such conversations, for example, with regard to coreferences. To this end, we build an alignment between the micropost conversation and some Semantic Web resources (Wikipedia Wikilinks, Wikidata, DBpedia). The solution presented in this article is one step in the constitution of the mirror graph, whose nodes are the items of the alignment (and edges are relationships taken from the Semantic Web resources used for building the alignment). The alignment includes other items than those resulting from the NEL step, that come from an Implicit Entity Linking (IEL) module. The NEL module produces a result from which the IEL one starts. One of our future work is to devise an incremental update of the mirror graph, this one becoming also an input to the EL process. The solution presented and evaluated in this article is generic and independent of such a configuration.

[6] https://fr.wikipedia.org/w/api.php
[7] https://www.wikidata.org/w/api.php

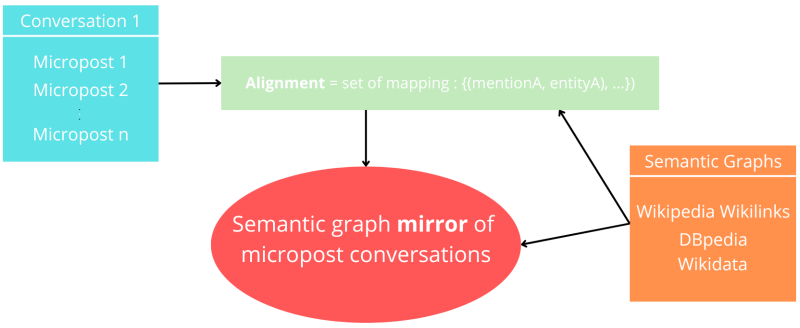

Fig. 2. View of the global framework

For now, the method (and its implementation) presented in this article is only used to build semantic graph mirrors of micropost conversations, for the CP4SC project. But it is very simple to understand, it is efficient, and it uses up-to-date Semantic Web resources that are accessible online without the need for huge downloads and processing. It is an independent brick, useful in any setting where semantic linking of mentions in microposts is necessary. It is freely made available to the community to answer the needs of more frugal solutions for such NLP tasks. Thanks to the high multilingualism of

The paper is organised as follows: in Section 2 we present related work for Entity Linking, Section 3 presents UFEL, our solution to link mentions in texts to entities in Wikidata, and Section 4 describes our experiments to asses it. We conclude in Section 5.

2 Related work

2.1 Named Entity Recognition

NER is a well-established task in natural language processing, but the majority of existing methods are designed for standard texts such as newspapers and extensive, well-organised documents. In the field's literature, first came the rule-based or symbolic approaches which depend on custom rules developed by experts. These methods are highly interpretable and may be effective for structured texts [7,23], but can not deal with microposts. There have been attempts to enhance them with machine learning, then came DL techniques with, for example, GRU-CRF models [24], which reaches 96.37% F1-score on the CoNLL-2000 dataset. However, nowadays some authors still suggest to use GPT-LLMs for this well-mastered task, for instance [32] employs a student-teacher architecture to train such a model in recognising entities within texts, in apparently complete ignorance of the far greater environmental impacts of these models on environment, as demonstrated in [3,16,17,4,29].

NER does not constitute our focus, but it is a preliminary step for the NEL task we are dealing with. Even if when applied to microposts, all NER systems face an important drop in performances [15], we chose to use one of the many pretrained basic DL solutions "on the shelf", without performing any new fine-tuning [29]. We acknowledge that even in this way, this still consumes a certain amount of resources (for future work, we will explore more resource-efficient solutions to align better with our goal of frugality), but for now we direct our attention to the core of our proposition, the NEL task.

2.2 Candidate generation

The initial phase of CG plays a crucial role as it involves creating a group of potential candidates associated with the entity mentioned. Two main approaches are used to generate candidates. First, graph-based strategies, e.g. [14], have proven their worth in generating contextually relevant candidates. These approaches typically involve traversing the KB graph structure (i.e. following connections between entities) to identify relevant entities based on predefined relationships. However, they depend on KB dumps that require significant storage, computational resources, and ongoing maintenance efforts to ensure their currency. In addition, they focus on a topological exploration of the KB to generate candidates, which can be limited and not adapted for our case due to the complexity of the data. In our case, where our objective is to select only a few highly relevant candidates, this approach might generate a large number of candidates, making the selection process less efficient. Second, DL-based methods [27,10,13] employ embeddings and similarity metrics to identify the most suitable candidates for a given mention. Although these approaches face similar constraints as those mentioned previously, they also require access to substantial amounts of training data.

To address these challenges, we have selected two resources that are freely usable. These two resources are the Wikipedia API[8] and the Wikidata API[9]. They are based on crowd-sourced KB, thereby resolving the update issue (i.e. keeping an up-to-date KB). Notice that by relying on KBs that evolve regularly, it may be more difficult to reproduce the experimental evaluations that demonstrate the relevance of UFEL. Since evolution consists in increasing and correcting the knowledge contained in these KBs, it is important to regularly report these evolutions in the annotations of the corpora used for evaluations, with a transparent corpus versioning policy, so that annotations can be updated transparently. Furthermore, these tools autonomously handle the search process, based on extensive indexation techniques. More importantly, they use various Wikipedia and Wikidata internal characteristics in their aggregation process, that are relevant to our situation, such as the frequency of node access, or the latest update of the node attributes. Both API are based on the mediawiki API [10] that allows to search through Wikipedia or Wikidata using different search

[8] `https://fr.wikipedia.org/w/api.php`
[9] `https://www.wikidata.org/w/api.php`
[10] `https://www.mediawiki.org/wiki/API:Search`

profile. There is 6 search profiles, each offering special features. For instance, for the *popular_inclinks_pv* profile, the entity ranking is based on the number of page views. These aspects of freshness, and diverse feature aggregations, align well with the data we are handling, as microposts conversations are most of the time about current affairs.

2.3 Candidate selection

The final phase of EL involves selecting for each mention the most appropriate entity from the list of candidates generated previously. Seminal works on candidate selection relied on textual similarity (e.g. [20]) to estimate semantic proximity. Such literal approaches have their limitation on complex cases where the consideration of contextual clues is really helpful. We take into account contextual information through the consideration of all couples of mentions and thus, all couples of related candidate sets.

Other contextual approaches have been considered in the literature. They involve graph-based methods that require the creation of a knowledge base to establish connections between graph nodes [11]. But their technique was restricted by the complexity of the identification problem. We were also inspired by TAGME [9], which used a semantic strategy to recognise the entity mentioned in short texts. However, it was based on a dump, and only considered the English Wikipedia as semantic resource. Lastly, the most sophisticated and powerful set of methods used for EL are DL techniques, which enable the abstraction of text content using embeddings that encompass contexts (e.g., [10]), facilitating CS based on candidates and their descriptions. These models have shown very good performances [31], but DL requires a large amount of training data and significant computational resources to be effective. In addition, neural systems are *black boxes* [5], preventing a complete understanding of the model. This is why we discarded DL approaches in favour of a heuristic algorithm that presents an appreciable transparency. It is based on multiple features, some of them computed from pre-trained embeddings.

3 UFEL: from mentions to entities

The solution we are proposing, that we call UFEL, is based on an analysis centred on the specific features of the texts we are dealing with. This is intended to optimise resource usage and to improve the clarity and understandability of the solution. It is founded on the idea that the entities mentioned successively in micropost conversations demonstrate strong semantic connections. This assumption is supported by the observation that entities in micropost data are closely linked within this context of the discourse. We tested this hypothesis on the annotated corpus we built to evaluate UFEL[11], by checking the shortest

[11] This resource is available here: `https://scm.univ-tours.fr/21202441t/corpus-ufel`. Codes for reproducing our experiments are also available here: `https://scm.univ-tours.fr/21202441t/ufel`

path existing between two consecutive entities behind two consecutive mentions. In 62% of the cases, the two entities are directly linked, by only one relationship within the Wikidata graph. We consider that this cohesion arises from two complementary reasons:

- Microposts are most often arguments to support opinions.
- Microposts are highly constrained by size. The resulting discourses are dense and focused on achieving the argumentative aim.

This phenomenon is particularly evident in textual forms such as microposts, but also in news articles, where the context evolves incrementally throughout the discourse building, with each mention contributing to the contextual framework and consequently expanding the set of entities within that context. This observation serves our objective of having a lightweight and flexible method, which is primarily motivated by two factors. First, we aim to provide a straightforward and replicable method that can be readily adjusted and enhanced. Second, the future industry demands such compact and effective methods that can be integrated into systems with minimal hardware and energy requirements.

As already explained, to focus on the NEL task, we selected a pre-trained NER model from the Hugging Face platform[12]. We evaluated multiple models on our corpus designed for the evaluation of the EL task, which will be described later. Table 1 presents the results of our experiments for the three best models among the ones we evaluated. We selected the solution with the highest recall, *roberta-large-NER*, in order to have as many mentions to link as possible.

Model	Recall	Precision	F1 score
Davlan-bert-base-multilingual-cased-ner-hrl	0.557	0.678	**0.653**
roberta-large-NER	**0.619**	0.692	0.624
bertweetfr-ner	0.604	**0.712**	0.613

Table 1. Results of three tested NER models, ordered by decreasing F1-score

3.1 Candidate Generation

The CG step is crucial, since if during this step we do not obtain the correct entities in the candidate set, there is no chance to link the mentions to their corresponding entity. For this step, we rely on external resources, considering an industrial context where storage is not always available. These resources are Wikipedia and Wikidata, together with their already mentioned APIs. Each of these resources offers the possibility, using a mention, to generate a set of Wikidata entities as candidates. They can use different search profiles, each using different available features (e.g. number of page access, page update, latest activity). As a result, each API returns a ranked list of entities based on the profile used for the querying process. After testing various profiles, the best

[12] `https://huggingface.co/`

results were obtained with the *engine_autoselect* profile. We use the two APIs because they sometimes return different results, and it is not always the same which returns the most relevant ones. We merge their results while checking whether they refer to the same entity, i.e. testing whether there is a semantic property *sameAs*[13] between them, when they differ.We estimated that in 85%-90% of the cases, the correct entity is within the first two candidates returned by both of these two APIs. The output of this step is CGr, a list of couples $[(mentionA, candidateSetA), \ldots]$, whose order is the order of mentions in the conversation.

Given the vast coverage of knowledge in Wikipedia and Wikidata, as can be seen from a simple search for a mention like "France"[14], simple search results can be very numerous, due to its great ambiguity when not contextualised. Leveraging Wikipedia and Wikidata APIs enables us to very significantly reduce the number of candidates retrieved, thereby streamlining the CS process. Simultaneously, this approach maximises the retrieval of correct entities, allowing to reach a balance between efficiency and precision in our system. This approach has multiple advantages. First, our method uses a contextual generation process that goes beyond surface form searches and uses multiple features to generate a coherent set of candidates. Second, this generation process does not rely on a KB dump, a type of approach that comes with complex problematics (i.e., freshness, storage, and querying performances). Third, microposts data and KB page access have similar activity in the sense that both are strongly influenced by current world events. For instance, during the Football World Cup, when querying the APIs for the mention "France", the entity behind the French National Team of Football will likely be ranked higher, which is coherent compared to social network data where, during the same period of time, "France" will more likely be associated with the French National Team of Football. With these two resources, the CG is robust and offers important retrieval performance.

Compared to the early NEL proposals based on ad hoc knowledge bases, significant time has elapsed and the Semantic Web technologies have enabled the emergence of very powerful tools based on valuable public resources. This enables us to rely now on those very rich public semantic resources, such as Wikidata and DBPedia and their related services, which are digital commons well managed and easily accessible, providing numerous opportunities for utilisation.

3.2 Candidate Selection

After generating a set of candidates for each mention, we get CGr, the list of couples $[(mentionA, candidateSetA), \ldots]$, whose order is the order of mentions in the conversation. The next task involves selecting the appropriate entity from each of the candidate sets. Based on the previously presented assumption that the entities mentioned successively in micropost conversations demonstrate strong semantic connections, we select the best entity in the set $candidateSetA$ of

[13] `https://www.w3.org/TR/2004/REC-owl-semantics-20040210/#owl_sameAs`
[14] Try it here: `https://www.wikidata.org/wiki/Special:Search`

Algorithm 1: Candidate Selection

Input: CGr : list of couples $[(mentionA, candidateSetA), \ldots]$
Output: SCC : list of triples $[(mentionA, entityA, scoreA), \ldots]$

```
1  SCC ← ∅; n ← card(CGr)
2  if n = 0 then
3      return SCC
4  if n = 1 then
5      return selectUnique(CGr)
6  bestTriple ← compareFirst(CGr)
7  SCC ← add(SCC, bestTriple)
8  for i = 2 to n do
9      bestTriple ← compare(CGr, i, SCC)
10     SCC ← add(SCC, bestTriple)
11 return SCC
```

candidates for $mentionA$ based on the set of candidates for the following mention in the list CGr, let's call it $mentionB$. This results in a list of selected candidates. More precisely, the result is $SCC = [(mentionA, entityA, scoreA), \ldots]$, a list of triples where each triple denotes the mention, its selected entity, and the score obtained by this entity in the selection process.

This is formalised in Algorithm 1: if CGr contains only one couple ($mention$, $candidateSet$) then Function $selectUnique$ selects the best entity by syntactic similarities between the mention and the labels of the candidate entities (lines 4-5). Otherwise, Function $compareFirst$ initialises the pair-wise comparison (lines 6-7), then Function $compare$ is used to process all couples in CGr (lines 8-10) before returning the resulting SCC.

The functions $compareFirst$ and $compare$ are based on the same logic, the only difference is their input. The initialisation ($compareFirst$) uses CGr to select the best entity for its first mention, by comparing the candidates of its first two mentions. To this end, it performs a cartesian product of their two candidate sets. For each item of this cartesian product, a similarity score is computed. The item with the highest score is selected: its first part becomes the best entity for the first mention of CGr, and is returned as the result. Function $compare$ takes as input CGr, the index i of the mention for which it must select the best entity, and the best entity previously selected for the preceding mention in CGr, which is contained in SCC. Function $compare$ returns the triple consisting of the ith mention, its selected entity, and the score of this entity. To this end, it computes a similarity score between the best entity e_{i-1} (previously selected for mention m_{i-1}) and all candidate entities for mention m_i, and selects the entity with the highest score among the candidates.

In addition to the successive-mentions hypothesis, the calculation of the similarity score between two entities, with their respective mention, is the kernel of our proposal. We tested numerous solutions before converging on the combination of measures that are presented in Algorithm 2. Each of the intermediate

Algorithm 2: Similarity Score of two entities with their mention

Input: $nbGraphNodes$: number of nodes in the KG, eA and eB: two entities, $mentionA$ the mention for eA, $mentionB$ the mention for eB
Output: The similarity score between eA and eB

1 $nbIncomingCA \leftarrow nbIncoming(eA)$
2 $nbIncomingCB \leftarrow nbIncoming(eB)$
3 $jaccard \leftarrow Jaccard(eA, eB)$
4 $shortestPath \leftarrow ShortestPath(eA, eB)$
5 $commonNodes \leftarrow CommonNodes(eA, eB)$
6 $cosimilarity \leftarrow Cosimilarity(eA, eB)$
7 $levenshteinA \leftarrow Levenshtein(mentionA, eA)$
8 $levenshteinB \leftarrow Levenshtein(mentionB, eB)$
9 return $scoring(nbGraphNodes, nbIncomingCA, nbIncomingCB, jaccard,$
10 $shortestPath, commonNode, cosimilarity, levenshteinA, levenshteinB)$

computations presented in Algorithm 2 are described in Table 2, where G is the semantic graph used to calculate the features.

Feature name	**Description**
$nbIncoming(e)$	Number of incoming links for the entity e in G
$Jaccard(e_1, e_2)$	Jaccard distance[12] for entities e_i types, based on G
$ShortestPath(e_1, e_2)$	Shortest path to go from entity e_1 to entity e_2 in G
$CommonNodes(e_1, e_2)$	Number of nodes in common in the neighbourhood of entities e_1 and e_2 in G
$Cosimilarity(e_1, e_2)$	The cosine similarity of e_1 and e_2 based on Wikipedia2Vec [30]
$Levenshtein(e, m)$	The Levenshtein distance between entity e's label and mention m

Table 2. Description of the variables used in Function *scoring*

The *scoring* function called in line 9 of Algorithm 2 is based on the following formula, that consists in in combining features that aim to capture different aspects of the selection task. Each of these features is presented in Table 3. Regarding the coefficients, we assigned the distinct weights to each component by a decision-making process, aiming to reflect the unique insights extracted from the data by each of the components. Using the hyperparameter optimisation tool Optuna [1] through its Python interface[15], we determined these weights through an optimisation process guided by the *Define by Run* principle. Optuna dynamically constructed the search space, ensuring a streamlined and robust approach to hyperparameter optimisation. The resulting values for UFEL are the

[15] `https://optuna.org/`

following ones: $\lambda_{SP} = 0.65$, $\lambda_{REL} = 0.7$, $\lambda_{JA} = 0.4$, $\lambda_{CO} = 0.9$, and $\lambda_{LEV} = 0.6$.

$$score(A, B) = \lambda_{SP} \times SP + \lambda_{REL} \times REL + \lambda_{JA} * JA + \lambda_{CO} \times CO + \lambda_{LEV} \times LEV \quad (1)$$

This scoring formula captures different features, which can all be important in the decision process when choosing between a set of candidates. This multiplicity of features and what they can extract is necessary to tackle the task of NEL when applied to microposts, due to the complexity of this type of data. The selection process is a complex task that needs to rely on a contextual approach, based on multiple entities, similar to what a human can do when it comes to disambiguating a mention during the reading process. Each feature used in the scoring formula aims at capturing a particular aspect of the existing relationship that can help select the correct entity.

Variable	Description
SP	The Shortest Path between the entities A and B
REL	The relatedness of A and B, based on [21]
JA	Jaccard's distance for A and B types
CO	The cosine similarity between the embedding vector for A and B, based on [30]
LEV	The Levenshtein distance between the two entity labels

Table 3. Features composing the scoring formula, with A and B two entities

4 Evaluation

4.1 Evaluation data

We performed many experiments to evaluate UFEL, *i.e.* the workflow: (i) selection of mentions resulting in a set of mentions and their position in the conversation, (ii) computation of the candidates resulting in CGr, and (iii) selection of entities that outputs SCC. We encountered a significant challenge for doing it, due to the scarcity of benchmarks. There is no recent and usable corpus of microposts in French, and even less that are correctly annotated for EL. There exists a corpus of tweets associated with the NEEL 2015 challenge, as detailed in Rizzo's work [25]. Unfortunately, our extensive efforts to access the original X data were unsuccessful. This corpus only comprises tweet IDs rather than actual tweet texts. The recent changes in X's API policy, which now demand payment for access, have made it even more challenging for us to access this dataset. The same problem occured with [22]. Therefore, we decided to create our own corpus. We provide it to reviewers so that they can judge UFEL's usability and the reproducibility of our experiments, but we still have to be sure that it is completely anonymised before we can open it to the general public.

Our corpus development was guided by the annotation guidelines used in the Impresso annotation campaign[8]. We focused on conversations related to

Conversations	568
Microposts	1,998
Token	47,646
Named Entity (NE)	2,781
Non-NIL link to entity	2,574
NIL	207

Table 4. Description of the corpus

cities (e.g. Paris, Bordeaux) to increase the volume. Our corpus includes 568 conversations, with a total of approximately 2,000 microposts. An excerpt of the corpus is shown in Figure 3. We restricted the extraction to subjects about cities, and we did not select any particular micropost type (e.g. formal, informal). The selected microposts were annotated for the EL task, which encompasses both the NER and the NEL components. More quantitative details are provided in Table 4. This newly compiled corpus serves as the basis for our evaluation process. Building it very carefully to be representative of French microposts conversations in general allows us to think that the results we obtain on this corpus can be generalised to any set of French microposts conversations.

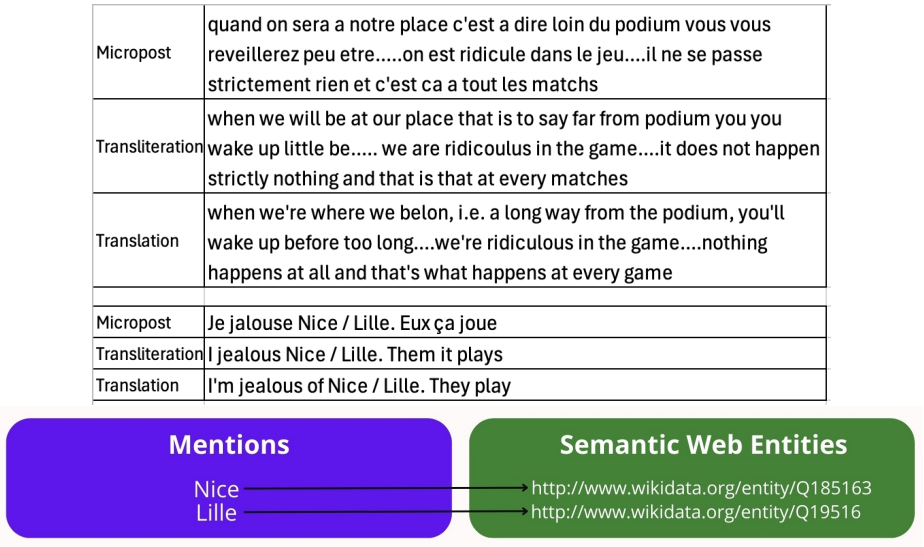

Fig. 3. Example of two microposts, their mentions and linkages

4.2 Protocol

We employed two distinct sets of measures: F1 score, recall, and precision on one hand, and alternatively, the metric outlined in [25], which is characterised as $score = 0.4 * mention\_ceaf + 0.3 * strong\_typed\_mention\_match + 0.3 * strong\_link\_match$. It offers an aggregative view of the whole process, considering the NER part and the NEL part. To assess the performance of UFEL, we deliberately opt for a diverse selection of other existing systems, acknowledging that our approach itself deviates from the mainstay of DL. Initially, we chose DBPedia Spotlight [19], a statistically oriented model grounded on DBPedia, accompanied by the Spacy Entity Linker, a string similarity-driven system comparing the recognised mention with the candidate's alias strings. We also tested Babelfy, a statistical system based on BabelNet, which formerly led the domain as State-Of-The-Art (SOTA) before the emergence of DL. Finally, ReFinED [28], a distinguished DL approach showing high performance, became our main reference point for comparisons. Each system serves as a robust reference point for comparison with the results of UFEL.

Using DBPedia Spotlight, Spacy Entity Linker, and Babelfy did not necessitate additional training, as these resources were pre-optimised for the French language. However, ReFinED required huge storage and computations due to the originally anglocentric orientation of the available models. Undertaking the French adaptation, we had to train a new model, which enabled us to clearly see the cost of such solutions in terms of computing and storage resources. First, we acquired and preprocessed the necessary files and then trained the model on three Nvidia Tesla V100 GPUs via the High Performance Computing (HPC) CaSciModOT cluster[16]. It took 40 hours of training and consumed approximately 36 KwH. For the training of the new French model, we strictly followed the instructions provided in the GitHub repository[17]. The results of the evaluation of UFEL are provided in Table 5.

4.3 Results

Given the diversity of architecture of the assessed systems, it was only to follow a *black box* paradigm evaluation: the evaluation concerns only the final outputs of the systems, without distinguishing between the NER and EL phases. The results provided in Table 5 demonstrate the relevance of our model. It offers indeed an appreciable level of performances (F1 = 0.752 and NEEL = 0.685) that is significantly higher to other systems. Such an observation was expected, since UFEL is devised specifically for micropost data, unlike other systems, which play here the role of reference baselines above all. However, we were surprised by the great sensitivity of these systems to the variation of language registers that represents micropost data. The repetition of the experiments confirms these conclusions: the performances observed remain very stable, with standard deviations of less than 0.05 for the metrics considered (Table 5).

[16] `https://cascimodot.fr/fr`
[17] `https://github.com/amazon-science/ReFinED`

This highlights the challenges involved in handling micropost data and underscores the need for a sophisticated approach to effectively capture the essential information required to disambiguate mentions.

System	P	R	F1	NEEL	Inference	CO2
UFEL	**0.751** ±0.02	**0.753** ±0.01	**0.752** ±0.03	**0.685** ±0.05	4min 9s	0.03
DBPedia Spotlight	0.196 ±0.001	0.378 ±0.001	0.234 ±0.001	0.269 ±0.001	10min 4s	-
Spacy EL[18]	0.41 ±0.001	0.158 ±0.001	0.228 ±0.001	0.4 ±0.001	6min 26s	-
Babelfy	0.264 ±0.001	0.330 ±0.001	0.293 ±0.001	0.219 ±0.001	7min 43s	-
ReFinED	0.313 ±0.04	0.272 ±0.05	0.291 ±0.04	0.381 ±0.03	7min 22s	0.87

Table 5. Performances, computing times, and carbon footprints (kg CO2 eq.)

A manual analysis of error has shown that the primary cause of UFEL's failure to connect a mention to the accurate entity is the absence of the correct entity in the set of generated candidates for that mention. The absence of a suitable candidate in the candidate pool is due to the generation process. When an entity no longer aligns with the fundamental assumption that the entity from the Wikidata or Wikipedia graph and the entity mentioned in the text share similar syntactic features, the generation process becomes inadequate.

Notice that Table 5 also includes running times and an estimation of carbon footprints (in kg CO2 eq.), calculated using the *TransparentToolkit's calculator* [18], accessible online[19]. It takes into account computing times, the type and the number of CPU/GPU used, the region of the world and the provider: in our case, OVHCloud is the most similar to the CaSciModOT cluster. As we do have access to all this information only for UFEL and RefinED, we provide numbers only for them. The comparison between UFEL and RefinED clearly shows the importance of designing a frugal approach such as UFEL. Remember that ReFinED also required almost 3 days for training, and that carbon emissions are just a small part of the environmental impact of digital technologies.

4.4 Notes on how Semantic Web could improve EL task's evaluation

NER evaluation When it comes to evaluating the complete EL pipeline, there are several limitations of the evaluation metrics used. First, when evaluating NER in the context of an EL pipeline, we consider a prediction to be correct when the full mentions have been captured by the system, that is, independently of the EL context. But only a part of the mentions can be sufficient to capture enough textual *grip* to later link it to the correct entry in the KB. For instance, taking the mention "Emmanuel Macron, the president of the republic", if the NER system captures only "Macron", the EL pipeline will certainly be capable of linking the identified span to the correct entry, i.e. *Emmanuel_Macron*. But

[19] `https://calculator.linkeddata.es/`

the evaluation only retains that the (full) mention is not in the result. This remark represents a great percentage of cases in which we evaluate the NER task without considering why we primarily use the NER. For EL, the NER is used to identify mentions inside a text, only to later link them to the correct KB entry. Thus, only the minimal text span that makes this linkage process possible is necessary. Even if, in a significant number of other use cases, the full span of NEs is necessary. Therefore, there should exist diverse evaluation protocols, in order to employ the appropriate one considering the contextual aim of the evaluation.

Candidate selection Second, when considering the evaluation of the selection candidate step, a similar problem arises. In the context of Semantic Web, we are dealing with entities highly inter-related with diverse relationships, and topologically speaking, groups can be formed. These groups representing strong relationship or similarity between their entities. Concretely, when evaluating using standard NLP metrics, such as the F1 score, we need to keep in mind that we are evaluating a potential *one-to-many* relationship with a *one-to-one* metric. For instance, when a system links the mention "Paris", which should be mapped to *https://dbpedia.org/page/Paris*, to *https://dbpedia.org/page/Ile-de-France*, this is a minor mistake given that *Île-de-France* is the region encompassing Paris, which means that they are strongly related. This direct link is present in the KB, thus it should be used. But on the basis of a one-to-one mapping, the result *https://dbpedia.org/page/Ile-de-France* is just noted wrong. For this example, in the context of semantic linking, it conveys a certain amount of quality, an information that can be relevant, and therefore we should not consider it as strictly wrong. The F1 score is still relevant in a significant number of cases, where only a one-to-one mapping exists. But in some cases, we should enhance the result obtained when it has a direct semantic link with the right entity. As a global remark, when using semantic Web resources and technologies, we can use more diverse metrics than a one-to-one syntactic mapping, for instance semantic distances existing between the gold entity and the one predicted by the system. Semantic Web resources can bring more diverse and meaningful evaluation strategies to the EL task.

5 Conclusion

We presented an approach that is light, fully understandable, and flexible to tackle the NEL task applied to micropost data, based on a simple observation of the data features on the one hand and, on the other hand, on the immense semantic *digital commons* that have appeared thanks to the Semantic Web principles, techniques, and tools over the last ten years. Their quality and representativeness continue to grow, thanks to the human communities that have embraced them. Relying on these remarkable joint efforts, our approach relies on transparent, fresh, verifiable and commonly understandable knowledge. Moreover, Wikipedia and Wikidata APIs are very efficient and easily usable services, using them requires no storage and very short computation times. This is super important for

the future of industry in the physically limited world we inhabit, which can be understood reading works such as [3,16,17,29]. As analysed in [4], digital technologies are a part of all scenari of future, but their costs and their uses will have to change. The digital commons are a good basis for this purpose, and the Semantic Web is a particularly relevant framework for the development and exploitation of those digital commons. The solution presented is intended to be part of a software suite developed by ATOS and dedicated to social network analysis, used for instance in the *Cloud Platform For Smart City* (CP4SC) project. Our experiments demonstrate satisfying performances while showing that SOTA DL solutions perform very disappointingly on French microposts. To perform these experiments, as no public benchmark was freely available for French microposts, we carefully devised a corpus, representative of French microposts conversations. We are currently conducting the complete pseudonymisation of this resource. It will be publicy released under a open licence when this process will be achieved.

UFEL could benefit from the integration of additional features, particularly by exploiting more entity relationships in Wikidata, for complementing the already used relatedness feature and potentially enhance the system's accuracy and contextual understanding. This is one of our future work, together with the general framework for constructing semantic graphs that mirror conversations of microposts, sketched in Introduction. Moreover, for now our experiments are specifically on microposts, more precisely on French microposts, which is necessary for our current objectives, but may result in a lack of generality. We plan to extent in the future their scope by adapting and assessing UFEL on other natural languages (English in particular) as well as on other social media communications. Additionally, we employ a deep learning-based solution for the Named Entity Recognition (NER) stage, which limits our aims of frugality and transparency. On this point, we plan to experiment with a finite-state transducer cascade solution that has already demonstrated high level performances on French spontaneous speech, another non-standard type of language. Finally, we do not address yet a well known limitation of pipeline architectures such as *Named Entity Recognition - Candidate Generation - Candidate Selection*, which is error propagation all along the pipeline of sub-tasks. We are currently launching a set of specific experiments to better characterise the potential influence of this propagation on the whole system's performance.

Supplemental Material Statement: Source code for UFEL is available at `https://scm.univ-tours.fr/21202441t/ufel`. The corpus used for evaluation is available at `https://scm.univ-tours.fr/21202441t/corpus-ufel`.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)

2. Antypas, D., Ushio, A., Barbieri, F., Neves, L., Rezaee, K., Espinosa-Anke, L., Pei, J., Camacho-Collados, J.: Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In: Findings of the Association for Computational Linguistics: EMNLP 2023 (2023)
3. Bannour, N., Ghannay, S., Névéol, A., Ligozat, A.: Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In: Moosavi, N.S., Gurevych, I., Fan, A., Wolf, T., Hou, Y., Marasovic, A., Ravi, S. (eds.) Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, SustaiNLP@EMNLP 2021, Virtual, November 10, 2021. pp. 11–21. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.SUSTAINLP-1.2, https://doi.org/10.18653/v1/2021.sustainlp-1.2
4. Bugeau, A., Ligozat, A.: How digital will the future be? analysis of prospective scenarios. CoRR **abs/2312.15948** (2023). https://doi.org/10.48550/ARXIV.2312.15948, https://doi.org/10.48550/arXiv.2312.15948
5. Delaunay, J., Galárraga, L., Largouët, C.: Does it make sense to explain a black box with another black box? CoRR **abs/2404.14943** (2024). https://doi.org/10.48550/ARXIV.2404.14943, https://doi.org/10.48550/arXiv.2404.14943
6. Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2), 32–49 (2015)
7. Díez Platas, M.L., Ros Munoz, S., González-Blanco, E., Ruiz Fabo, P., Alvarez Mellado, E.: Medieval spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. Journal of the Association for Information Science and Technology **72**(2), 224–238 (2021)
8. Ehrmann, M., Watter, C., Romanello, M., Simon, C., Flückiger, A.: Impresso named entity annotation guidelines (clef-hipe-2020). Tech. rep. (2020)
9. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1625–1628 (2010)
10. Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. arXiv preprint arXiv:1704.04920 (2017)
11. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 765–774 (2011)
12. Jaccard, P.: Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. Bull Soc Vaudoise Sci Nat **37**, 241–272 (1901)
13. Le, P., Titov, I.: Improving entity linking by modeling latent relations between mentions. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018)
14. Li, D., Fu, Z., Zheng, Z.: An entity linking model based on candidate features. Social network analysis and mining **11**(1), 50 (2021)
15. Liu, P., Wang, G., Li, H., Liu, J., Ren, Y., Zhu, H., Sun, L.: Multi-granularity cross-modal representation learning for named entity recognition on social media. Information Processing and Management **61** (1 2024). https://doi.org/10.1016/j.ipm.2023.103546
16. Luccioni, A.S., Hernández-García, A.: Counting carbon: A survey of factors influencing the emissions of machine learning. CoRR **abs/2302.08476** (2023).

https://doi.org/10.48550/ARXIV.2302.08476, https://doi.org/10.48550/arXiv.2302.08476

17. Luccioni, A.S., Viguier, S., Ligozat, A.: Estimating the carbon footprint of bloom, a 176b parameter language model. J. Mach. Learn. Res. **24**, 253:1–253:15 (2023), http://jmlr.org/papers/v24/23-0069.html
18. Markovic, M., Garijo, D., Germano, S., Naja, I.: Tec: Transparent emissions calculation toolkit. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., Li, J. (eds.) The Semantic Web – ISWC 2023. pp. 76–93. Springer Nature Switzerland, Cham (2023)
19. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8 (2011)
20. Mihalcea, R., Csomai, A.: Wikify! linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242 (2007)
21. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 509–518 (2008)
22. Mishra, S., Saini, A., Makki, R., Mehta, S., Haghighi, A., Mollahosseini, A.: Tweetnerd-end to end entity linking benchmark for tweets. Advances in Neural Information Processing Systems **35**, 1419–1433 (2022)
23. Moncla, L., Gaio, M., Joliveau, T., Lay, Y.F.L.: Automated geoparsing of paris street names in 19th century novels. In: Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities. pp. 1–8 (2017)
24. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108 (2017)
25. Rizzo, G., Basave, A.E.C., Pereira, B., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.: Making sense of microposts (# microposts2015) named entity recognition and linking (neel) challenge. In: # MSM. pp. 44–53 (2015)
26. Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., Biemann, C.: Neural entity linking: A survey of models based on deep learning. Semantic Web **13**(3), 527–570 (2022)
27. Sun, Y., Lin, L., Tang, D., Yang, N., Ji, H., Wang, X.: Modeling mention, context and entity with neural networks for entity disambiguation. In: Proceedings of the 24th International Conference on Artificial Intelligence (2015)
28. Tom Ayoola, Shubhi Tyagi, J.F.C.C.A.P.: ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In: NAACL (2022)
29. Wang, X., Na, C., Strubell, E., Friedler, S., Luccioni, S.: Energy and carbon considerations of fine-tuning BERT. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023. pp. 9058–9069. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.607, https://doi.org/10.18653/v1/2023.findings-emnlp.607
30. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. arXiv preprint arXiv:1812.06280 (2018)
31. Yamada, I., Washio, K., Shindo, H., Matsumoto, Y.: Global entity disambiguation with bert (2022)

32. Zhou, W., Zhang, S., Gu, Y., Chen, M., Poon, H.: Universalner: Targeted distillation from large language models for open named entity recognition. arXiv preprint arXiv:2308.03279 (2023)