Indirectly Named Entity Recognition

Alexis Kauffmann, François-C. Rey, Iana Atanassova, Arnaud Gaudinat, Peter Greenfield, Hélène Madinier, Sylviane Cardey

Kauffmann, Alexis, Rey, François-Claude, Atanassova, Iana, Gaudinat, Arnaud, Greenfield, Peter, Madinier, Hélène, and Cardey, Sylviane (2021) Indirectly Named Entity Recognition. Journal of Computer-Assisted Linguistic Research 5: 27-46. ISSN: 2530-9455.

Indirectly Named Entity Recognition

Alexis Kauffmann*[1], François-Claude Rey[2], Iana Atanassova[2,3], Arnaud Gaudinat[1], Peter Greenfield[2,4], Hélène Madinier[1], Sylviane Cardey[2,3]

[1] HES-SO/HEG Genève, Switzerland

[2] CRIT, University of Bourgogne - Franche-Comté, France

[3] Institut Universitaire de France (IUF), France

[4] Peter Greenfield passed away during the preparation of this manuscript.

* Corresponding author: alex_kauffmann@yahoo.fr

Received: 14 July 2021 / Accepted: 21 October 2021 / Published: 22 December 2021

Abstract

We define here *indirectly named entities*, as a term to denote multiword expressions referring to known *named entities* by means of periphrasis. While named entity recognition is a classical task in natural language processing, little attention has been paid to *indirectly named entities* and their treatment. In this paper, we try to address this gap, describing issues related to the detection and understanding of *indirectly named entities* in texts. We introduce a proof of concept for retrieving both lexicalised and non-lexicalised indirectly named entities in French texts. We also show example cases where this proof of concept is applied, and discuss future perspectives. We have initiated the creation of a first lexicon of 712 indirectly named entity entries that is available for future research.

Keywords: named entities, indirectly named entities, information extraction, named entity recognition, multiword expressions, text processing, text mining

1. Introduction

We define here *indirectly named entities* (INE) as multiword expressions (as defined in Baldwin and Kim 2013) referring to known *named entities* by means of periphrasis (as in Racicot 2009), such as *the City of Light* for Paris or *the Big Apple* for New York. The use of INE in a text, instead of explicitly using the related named entities, requires the reader to detect that a periphrasis is applied and to understand, or guess, what named entity the periphrasis refers to. These two issues, detection and understanding of INE, we consider to be suitable tasks to be addressed as Natural Language Processing and Information Retrieval. They are discussed in this paper.

While much research has been done on Named Entity Recognition (NER), the problem of

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2021.15922
PUBLISHER: UNIVERSITAT POLITECNICA DE VALENCIA.

processing INE in texts to our knowledge has not been addressed yet. For applications such as information retrieval, anonymisation or semantic analysis, finding INE can be useful. This task, which we name here *Indirectly Named Entity Recognition* (INER), can be seen as a complement to classical NER, which is inefficient for detecting most of INE.

The rest of the paper is organised as follows: In the first section, we present the state of the art of existing natural language processing and information retrieval work that address this phenomenon or related cases, we describe the linguistic phenomenon of INE, and we mention closely related cases. In Section 2, we propose a model for INER which is based on both lexical data and also linguistic rules. In Section 3, we explain how we have implemented this model and show example cases of our model performing on French text. In the final section, we discuss further work, such as understanding INE and finding the named entity hidden behind an INE, especially in ambiguous or unknown cases.

1.1. Related work

To our knowledge, INER has not been studied yet, but is related to several other domains in Natural Language Processing (NLP), in particular multiword expressions, NER, anaphora resolution and entity linking, and paraphrase identification.

1.1.1. Multiword expressions

Multiword expressions (MWE) have been the object of numerous studies in NLP (Baldwin and Kim, 2013). For example, Wehrli et al. (2010) have studied collocations (mainly verb-object collocations), and how to detect and translate them, taking advantage of multiword lexical data, lexical tagging and syntactic parsing (see also Wehrli and Nerima 2018; Kauffmann 2013). Our model for INER is partly inspired by these works.

1.1.2. NER

NER enables detecting named entities in a text (Friburger, 2006; Nadeau and Sekine, 2007; Lopez et al., 2019). This task has been achieved following rule-based and hybrid methods (Charton et al. 2011; Lopez et al., 2019), or deep neural models (Lample et al. 2016; Shang et al. 2018; Lin et al., 2020; Zhang et al., 2020). For the evaluation of NER in French, Ortiz Suarez et al (2020) have created an annotated version of the French TreeBank corpus, and several methods for NER evaluation have been discussed (Nouvel et al., 2016). Moreover, NLP software able to achieve NER making use of deep neural methods is now publicly available (Schmitt et al., 2019). Among these, we find Bert (Devlin et al., 2019; Tenney et al., 2019), CamemBert (Martin et al., 2020), Spacy (Honnibal and Montani, 2017) and Stanford CoreNLP (Manning et al., 2014).

The question of retrieving multiword named entities that may not have been lexicalised yet or that may have a different form compared to the lexicalised ones, is addressed in Nayel et al. (2019) for the biomedical domain and in Watanabe et al. (2019) for the chemical domain. Our study here also aims at showing how to detect INE that may not have been lexicalised and that may not be detected by traditional NER tools, but without focusing on specific domains such as the biomedical or chemical ones.

1.1.3. Anaphora resolution

As INE always refer to named entities, pronouns refer to referential lexemes or named entities in texts. Anaphora resolution (or coreference resolution) aims at solving anaphora and linking, for example, pronouns with the named entities they refer to (Mitkov, 2014; Ma et al., 2020). Coreference and anaphora resolution for French have been studied in Guenoune et al. (2020).

1.1.4. Entity linking and synonym entity recognition (SER)

Named entity synonyms have been the object of several studies that define the task of Named Entity Normalisation (NEN), e.g. in the biomedical domain (Cho et al., 2017) or in materials science (Weston et al., 2019). In a broader context, synonym entity recognition aims at detecting synonymous named entities in a text, such as, for instance, *Hollande* and *François Hollande* (Ananthanarayanan et al., 2008; Bohn and Nørvag 2010; Chakrabarti et al., 2012; Qu et al., 2017; Cai and Wu 2019; Shen et al., 2019; Yang et al., 2021). This task plays an important role in many applications such as information retrieval, information extraction and question answering.

The question of linking INE or other synonym entities with the named entities they refer to is addressed in Rosales-Mendez et al. (2019) and Wu et al. (2018) and described as "entity linking", but this research does not specifically deal with INE recognition, as we do here. Disambiguation among several homonym candidates in entity linking is addressed in Nebhi (2013), Qu et al. (2017) and Wu et al. (2018).

1.1.5 Paraphrase identification

Synonym entity recognition has also been evocated in the more global context of "paraphrase identification" (Mohamed and Oussalah, 2020; Wanachai and Cardey-Greenfield, 2012; Shinyama et al., 2002, Sales et al., 2016). In this context, possible variations in the expression of phrases or whole sentences are studied, and the use of INE or other synonym entities appear as possible aspects that these variations can take.

Compared to the all related works that we have found, our study is innovative because it focuses specifically on INE and their morpho-syntactic, lexical and semantic properties, and, making use of these properties, defines a model for INER. INER can be useful as a complement to NER, since INE are not necessarily detected by NER tools.

1.2 Indirectly named entities: Definition and linguistic study

As mentioned earlier, we define INE as multiword expressions referring to known *named entities* by means of periphrasis: INE, instead of mentioning the usual name of the named entity, refer to it in an implicit and descriptive way. They do not constitute an alternative name but rather a periphrastic description of the named entity, which is specific enough to be lexicalised.

Periphrases can be made on named entities in any human language. However, some of their syntactic and lexical properties can be language-specific. We will focus here firstly on French and then English.

(1) *la ville lumière* (*Paris*)
the city (of) light

(2) *la cité de Calvin* (*Geneva*)
the city of Calvin

(3) *la Venise des Alpes* (*Annecy*)
the Venice of the Alps

(4) *la petite Venise des Alpes* (*Annecy*)
the little Venice of the Alps

(5) *le président des riches* (*Nicolas Sarkozy* or *Emmanuel Macron*)
the president of the rich people

(6) *le Président des riches* (*Nicolas Sarkozy* or *Emmanuel Macron*)
the president of the rich people

(7) *l'Hexagone* (*France*)
the hexagon

(8) *outre-Manche* (*Great Britain*)
the other side of the Channel

In French, we have listed three frequent syntactic structures of INE. The first typical syntactic structure is a determiner phrase[1] which, unlike most French proper nouns, always start with a definite article (*le*, *la* and their derived forms), as in examples 1 to 6), and which contain a head noun and one noun modifier (such as an adjective, a relative clause or a prepositional phrase), typically resulting in a "DET-N-MOD"[2] structure or a "DET-MOD-N" structure. Sometimes, the determiner phrase contains more than one noun modifier (as in ex. 4). In (Racicot, 2009) are described famous French and English periphrases applied to geographical entities, such as examples 1 and 12, and we can observe that the "DET-N-MOD" and "DET-MOD-N" structures cover the majority of examples.

The second typical syntactic structure of INE is a simple "DET-N" structure[3], as in example 7.

The third typical syntactic structure of INE, less frequent, (as in example 8), consists of an adverbial phrase made by adjoining a prepositional prefix to a named entity.

We can see that, unlike French named entities and proper nouns, the head nouns of French INE do not necessarily start with a capital letter. The use of a capital letter for the head noun is

[1] We will consider here that the determiner is the syntactic head of the noun phrase, even if the noun is the semantic head, following the DP hypothesis (Abney, 1987). So, we can call it DP (Determiner Phrase), even for cases where there is no determiner before the noun, which we consider a null determiner.

[2] Here DET means determiner, N means noun and MOD means modifier.

[3] This type of INE requires more contextual or cultural knowledge to be understood, since no modifier is applied to the head noun.

generalised only when the head noun is a proper noun (ex. 3). When the head noun is a common noun, some writers choose to capitalise its first letter (ex. 6) while, most often, it is not capitalised (ex. 5, 1, 2).

For English we have:

(9) The City of Light (Paris)

(10) The Five Boroughs (New York)

(11) The Big Apple (New York)

(12) Shaky Town (San Francisco)

(13) Big Blue (IBM)

(14) The Boss (Bruce Springsteen)

In English, as in most French cases, INE have the "DET-N-MOD" or "DET-MOD-N" structure (examples 9 to 13), or a "DET-N" structure, as in example 14. However, a null determiner is sometimes found instead of the definite article, as in examples 12 and 13.

In English, in lexicalised periphrases on named entities, the nouns, adjectives and adverbs usually start with capital letters (as in examples 9 to 14).

1.3 Related cases

Here we discuss three different cases: synonymous named entities, non-implicit nicknames, and generic anaphora, which are related to INE, but are different.

1.3.1 Related cases: Synonymous named entities

(15) Pusan (Busan[4])

(16) St-Étienne (Saint-Étienne)

(17) The UN (The United Nations)

Synonymous named entities such as alternative spellings (ex.15), abbreviations (ex.16), or acronyms (ex.17) of named entities do not match our definition of INE. Unlike INE, synonym named entities do not imply the identity of a named entity but explicitly name it, in a way that differs from the one which is usually considered as standard[5].

[4] The Korean city of Busan has its name originally spelled 부 산 in Korean. It is transcribed in the Latin alphabet transcribed as *Busan* in the Revised Romanization of Korean, which is the official Romanisation system in South Korea since 2000, and as *Pusan* in the McCune–Reischauer Romanisation system.

[5] We assume here that there is a de facto standard spelling or naming for a named entity, while we might not know, among synonymous named entities, which one would appear to be the de facto standard. The 'standard' can also change depending on the context. For example, French singer Johnny Hallyday was referred to as Johnny Hallyday in the media and music context while his official name, reference only in an administrative context, was Jean-Philippe Smet.

1.3.2 Related cases: non-implicit nicknames

INE can be classified as a subset of "aliases" (as in Rosales-Mendez et al. 2019) or "nicknames" (as in Racicot 2009). For instance, in Wikidata (and Wikipedia), where named entities are listed, there is no specific property field for their periphrases, but there is a "Nickname" property field (P1449 in Wikidata classification). The subset idea that *all INE can be seen as nicknames but not all nicknames are INE* is confirmed by inspecting examples of nicknames contained in Wikidata, especially personal and geographical nicknames: some nicknames are INE (similar to the examples given in Section 1.2), while most other are not.

(18) Chuck (Charles E. Tucker, Jr.)

(19) Fat Mike (Mike Burkett)

(20) Tom (Robert Dick Hutchins)

We call here *non-implicit nicknames* those nicknames that, as any typically named entity, contain a name and express it, instead of making use of a periphrasis. Among these, many are just a first name (as ex. 18 and 20), a name with an adjective or a descriptive element (ex. 19), or a different name for a same-named entity (ex. 20). Thus, non-implicit nicknames can rather be seen as informal synonymous named entities.

1.3.3 Other related cases: Generic anaphora

(21) *Le gardien du PSG a déclaré : "C'est une occasion en or pour notre équipe."*
The PSG goalkeeper has declared: "It is a golden opportunity for our team."

In the definition of INE we do not include generic and non-lexicalisable anaphora as shown in example 21. In this example, *notre équipe* (*our team*) refers to the PSG team, but this anaphora is nonspecific and entirely context-dependent: *notre équipe* (*our team*) may be used, in other contexts, to refer to *any* other team. Consequently, it is not appropriate for lexicalisation and is not a case of INE.

2. INER: Methodology for the identification of INE in texts

In this section, we propose a general methodology for the retrieval and identification of INE from texts. We discuss firstly the possibility of a simple lexical approach and then introduce our method that relies on a database of patterns of INE which is used together with a rule-based approach.

2.1 A lexical approach

Following the (Baldwin and Kim, 2013) classification of multiword expressions, INE can be fixed or semi-fixed expressions (see also Alsharaf et al., 2003). A basic approach for INER can be a purely lexical approach, where all possible INE are listed in a lexicon and retrieved by simple string matching in texts. A possible data structure for such a lexicon is a set of pairs in the form

("INE ; NE"[6]). The major limitation of such an approach is the fact that the lexicon needs to be designed to cover the largest possible set of INE, and INE that were not initially included in the lexicon cannot be retrieved. On the other hand, the existence of such a lexicon allows knowing what named entity is related to an INE, as well as the detection of ambiguous cases where several named entities are possible candidates for one INE.

As named entities can be semantically classified, so can be INE. The three fundamental semantic categories usually used for named entities, which we study here, are a) "locations", b) "persons", and c) "companies and organisations"; to which we have added a fourth fundamental semantic category which is d) "products and services" that encompasses also information systems and art. We have listed typical lexical-syntactic patterns for each of these four semantic categories, and detailed the types of INE that are included as follows:

- Locations: natural and artificial places and areas in space, physical geographical entities and territorial entities (countries, regions, towns, districts, etc.), natural and human paths and routes (rivers and deltas, streets and crossroads, etc.). E.g., *the country of Uncle Sam* for *the USA*, or *the most beautiful avenue of the world* for *the Champs Elysées* avenue in Paris. Are excluded: locations in time (periods, dates, etc., e.g. *the Great War*).

- Persons: individual human beings and real or fictional characters, personified pet animals, and personified objects (dolls, etc., with a given name). E.g., *the Iron Lady* for *Margaret Hilda Thatcher* as (former) Prime Minister of the UK, or *the King* for *Elvis Aaron Presley* as a singer.

- Companies and organisations: legal entities and *de facto* entities (commercial companies, administrations, associations, foundations, political parties, religious groups, etc.). E.g., *the Redmond's Giant* for *Microsoft Corporation*, or *Auntie* for the *British Broadcasting Corporation* (BBC). Are excluded: territorial entities (see the above *Locations* category definition).

- Products and services: hand-made or industrial-made or intellectually elaborated commercial materials and objects and actions, extracted raw materials and commodities, on sale or free of charge offered individually or collectively hold goods and possessions, information documents and materials and media, machines and technical systems, inventions, logos and trademarks, symbols and artistic works, natural languages. E.g., *the Beetle* for *the Volkswagen Type 1* car, or *white gold* for the *latex extracted from hevea bra*silie*nsis t*rees, or *the language of Pouchkine* for *Russian*.

2.2 Building a database of INE patterns and their named entity equivalents

As a generalisation of the lexical approach, we have constructed a dictionary of patterns of INE with their named entity equivalents. We define a pattern of INE as a linear sequence of surface linguistic elements (or linguistic expressions), where each linguistic element can be an item of a predefined list, or a linguistic expression (e.g. a word) having some specific formal characteristics that can be easily identified (e.g. a word starting with a capital letter and composed only of capital and lower-case letters).

[6] Here, INE stands for indirectly named entity and NE stands for named entity.

The following example shows a pattern of the type "companies and organisations". The named entity is the name of a company behind a brand, and is replaced by a description of the logo of the company:

- {<"la">(<" ">)<"marque"><" "><"au">(<" ">)<logo description>}

In this pattern, the <logo description> tag is representing a part of an INE, and the elements between the parenthesis are optional. The different tags represent lists of expressions, e.g. <*la*> (definite article *the*) can be replaced by another definite article; <*marque*> (noun *brand*) can be replaced by equivalent alternatives such as *firme*, *compagnie*, *établissement*, *entreprise*, *societé*, etc.; and <*au*> (preposition *with*) can be replaced by its plural *aux* or by its feminine *à la*.

Here is a match of the previous pattern in a real sentence: "Gilles Vidal, patron du Design de la marque au lion, vient de quitter ses fonctions.", where *la marque au lion* (*the lion brand*) stands for *Peugeot*. Amongst other examples are: *la marque à la couronne* (*the brand with the crown*) for *Rolex*; *la marque au Swoosh* (*the brand with the swoosh*) for *Nike*; *La marque à la petite bretelle rouge* (*the brand with the little red strap*) for *Caracoteen*.

The process for building the database of INE patterns follows three stages as described below.

1) Lists of INE are manually retrieved from texts and classified. For example, *la langue de Pouchkine* (*the language of Pouchkine)* can be found in texts written more than a century ago, e.g., in a french novel[7] ("[...] par quelques phrases prononcées dans la langue de Pouchkine.") and in its translation in English[8] ("[...] with some phrases pronounced in the language of Pouchkine."); and it can be found in recent texts, e.g., in a printed newspaper[9] ("La langue de Pouchkine reste largement parlée") and in English on a commercial Internet website[10] ("There are many different styles of writing in the language of Pouchkine:"). These lists are enriched with INE inspired by the work of (Bachellier 1972; Petit 2006; Friburger 2006; Racicot 2009; Treps 2012; Sjöblom 2016), as well as INE that are found in Wikipedia pages or newspaper articles.

2) The lists of INE are organized in subcategories according to their linguistic similarities and named entity equivalents. For example, the "person" category includes amongst other the following subcategories:

- Religious names or titles: e.g. a religious character title like *the Savior* for *Jesus of Nazareth*, and *the Mahatma* for *Mohandas Gandhi*.

- Political and mass media nicknames, e.g. *l'inconnu le plus célèbre de France* (*France's most famous unknown person*) for *Edouard Balladur* (former Prime Minister of France), and *le grand*

[7] In Klaczko Julian, "Deux chanceliers : le prince Gortchakof et le prince de Bismarck", 2nd ed., p. 438. E. Plon et cie, Paris, France, 1876. Text available online on https://gallica.bnf.fr.

[8] Klaczko Julian, and WARD Frank P. (translator), "Two chancellors: Prince Gortchakof and Prince Bismarck", p. 103. Hurd & Houghton, Cambridge, Massachusetts, USA, 1876. Text available online on https://archive.org.

[9] In Daum Pierre, article "En Géorgie, l'obsession de la Russie". Monthly newspaper *Le Monde diplomatique*, October 2021, n° 811, p. 4. SA Le Monde diplomatique, Paris, France.

[10] In article "Our Advice for learning Russian". Page of commercial website blog. Dated from the 18/02/2018. Consulted on the 12/10/2021 at https://www.superprof.co.uk/blog/start-learning-russian/.

Charles (*the great Charles*) for *Charles De Gaulle* (former President of France).

- Nicknames of (quasi-)official functions or duties, e.g. *le Pacha* for the commanding officer of a French warship or *le mandarin* for the chief medical officer of a French medical department, and *le grand timonier* (*the Great Helmsman*) for *Mao Zedong* (former President of China).

3) The patterns of INE are constructed by systematic observation of the lists for each of the subcategories. For each pattern, we also observe the inner and outer context of the INE in real sentences, and describe its signification.

The example below shows a more complex pattern from the category "locations" which is explained with two calculable patterns (pat1 and pat2) and one list of other alternative patterns (listPatAlt). This pattern matches INE corresponding to the named entities of each US state, e.g. *le 42ème état des États-Unis d'Amérique* (*the 42th state of the USA*) for *l'État du Washington* (*the State of Washington*):

- pat1 = {<listArtS><listN><listTH><" "><listSt>(<" des "><listUS>)
  (<" d'Amérique">(<" du Nord">))}
- pat2 = {<listArtS><listN><listTH><" "><listSt>(<" nord-">)<"américain">}
- listArtS = "Le ","le "
- listNfig = "1" to "50" (US states numbers in figures)
- listNlet = "premi","deux","trois","quatr","cinqu" to "cinquante-et-un"
- listN = listNfig + listNlet
- listTH = "er ","e ","eme ","ème ","ième "
- listSt = " état"," Etat"
- listUS = "EU","EUA","USA","U.S.A.","Etats-Unis","États-Unis"'
- listPatAlt = "l'état vert","l'État vert","l'Etat vert", "L'état vert", ...("the Evergreen State")

As a result of this manual database creation phase, we have obtained a list of patterns for each of the four fundamental semantic categories, where the information for each pattern consists of:

- the description of the pattern;
- the lists of (sometimes non-exhaustive) alternative expressions fitting some of the pattern parts in the paradigmatic axis;
- a list of INE (and corresponding *named entities*) that match the pattern;
- a (possible) list of exceptions to which the pattern should match;
- example sentences containing occurrences of the pattern (for future reference, testing, and subsequent automated benchmarking).

As a final result we have obtained 40 INE patterns: 8 for locations, 9 for persons, 9 for companies and organisations, and 14 for products and services. Each one of these 40 patterns, and of their variants when they have such, is registered with at least one corresponding

documented real sentence containing a matching expression.

2.3 Our model for INER: A lexical-based and rule-based approach

While the database of INE patterns described above allows the retrieval of a large number of occurrences in texts, such an approach is still dependent on the number and variety of patterns that are manually created. To tackle the problem of detecting non lexicalised INE and detecting neology in INE, we need a more general approach. We propose to use linguistic rules that are obtained after lexical tagging, syntactic parsing, NER and semantic tagging of named entities. In a similar approach, Wehrli et al. (2010) take advantage of the syntactic parsing for the detection of multiword expressions, even in derived cases where their lexemes are not strictly adjacent.

As an example, we can consider the INE *la Venise des Alpes* (ex.3). Various derived periphrases can exist such as *la petite Venise des Alpes* (ex.4), which is formed by the attachment of a new qualifier to the head noun. In this case, the typical structure of INE "DET-N-MOD1", is modified to "DET-MOD2-N-MOD1". The rules of such modifications can be obtained by the observation of the INE's possible periphrases. Their implementation requires several pre-processing steps that include the morpho-syntactic parsing of sentences.

For example, defining a detection rule for periphrases with a "DET-NE-P-GNE"[11] structure, with the syntactic structure defined in the syntactic tree (see figure 1), we can retrieve all the periphrases such as *la Venise des Alpes* (ex.3) but also any neologism matching with this rule, such as, for example, *the Brad Pitt of Egypt* or *the Siberia of Scotland*.

[11] Here and in the syntactic tree in figure 1, NE stands for named entity, GNE stands for geographical named entity, P for preposition, DET for determiner, PP for prepositional phrase, MOD for modifier, NP for noun phrase and DP for determiner phrase.

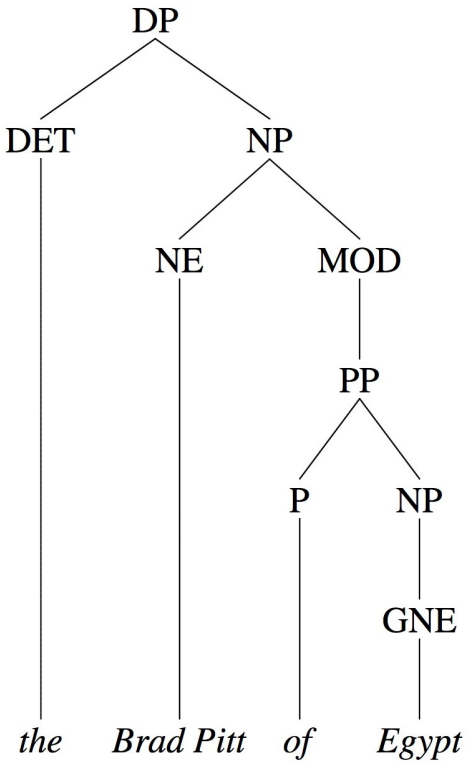

FIGURE 1. SYNTACTIC TREE FOR A DET-NE-P-GNE STRUCTURE AND EXAMPLE OF MATCHING INE

Thus, such syntactico-semantic, non lexicalised rules for the detection of INE can work together with the pattern-based retrieval and classification of INE, and allow the retrieving of even full neology in the INE that match with the defined rules.

3. Illustration of the concept

3.1 Implementation

As a baseline, we have built a lexical INER system, making use of the Spacy NER tool[12] and adding a lexical component to it, in Python, making use of the Spacy API. We were then able to do NER and INER in a single text analysis, in either French or English.

Then, in order to achieve lexical-based INER for French, we populated the lexicon with 878 French INE. We incorporated in this lexicon both data collected from diverse sources such as (Racicot, 2009) and classified as described in Section 2.1 and 2.2, and data found in crowd-sourced Wikipedia pages about geographical periphrases[13]. We have made this compiled lexical dataset publicly available on the open science platform Zenodo (Rey and Kauffmann, 2021). Some lines of the dataset are shown in Figure 2.

[12] https://spacy.io/usage/linguistic-features#named-entities

[13] https://fr.wikipedia.org/wiki/Liste_de_périphrases_désignant_des_villes and https://fr.wikipedia.org/wiki/Liste_de_périphrases_désignant_des_pays

la marque au vichy rose	Tati (entreprise)	ORG	marque
la marque au lion	Peugeot (entreprise)	ORG	marque
la marque au double chevron	Citroën (entreprise)	ORG	marque
la marque au taureau	Lamborghini (entreprise)	ORG	marque

FIGURE 2. ENTRIES FROM OUR FRENCH INE DATASET.

The columns are the following: INE, related NE (with additional information between parentheses), semantic category, subcategory.

Finally, for the lexical-based and rule-based approach, we implemented lexico-syntactic and syntactico-semantic detection rules, on the basis of the lexico-syntactic patterns that we have described in Section 2.2. We used for this task the Spacy lexical tagger and syntactic parser. The lexico-syntactic rules are directly inspired from the lexico-syntactic patterns of Section 2.2 and also make use of the Spacy "matcher" tool[14], while the syntactico-semantic rules make use of the Spacy NER tool with its semantic tags and are directly coded in Python.

3.2 Examples of retrieved INE

Since implementation and lexical work progressed further for French than English, we mention here only work on French.

As a baseline, we first tested the Spacy NER tool[15] and observed an evolution of the behaviour with some INE between September 2020 (Spacy 2.3.0) and March 2021 (Spacy 3.0.5). In version 3.0.5, some INE with a D-NE-P-NE structure are now detected as a whole entity in the NER task[16], while two independent named entities were detected in the 2.3.0 version. For example, *la Venise des Alpes* is detected as *Venise des Alpes* instead of detecting separately *Venise* and *Alpes* as in version 2.3.0.

Then, using a compiled INE lexicon and rules in addition to the Spacy 3.0.5 NER component, we have achieved NER+INER, on an example text (fully displayed in Figures 2 and 3) that contains 23 sentences, 6 named entities and 21 INE.

We can see that, in this example text, our method enables retrieving INE thanks to rule-based detection or the use of the INE lexicon.

Rule-based matching with the corresponding patterns enables retrieving the INE that have not been detected at the NER step: *la capitale du skate* (*the capital city of skateboarding*), *le Bratt Pitt*

[14] https://explosion.ai/demos/matcher

[15] In this case, without any specific retraining, using the model trained on the "fr core news sm" corpus.

[16] In this case, without including the initial definite article of the INE.

caché des Açores (*the hidden Brad Pitt of the Azores*), *le président des riches* (*the president of the rich people:* Nicolas Sarkozy or Emmanuel Macron), *Le pays des érables (the country of maple trees*: Canada*), la ville du bout du lac (the city at the end of the lake*: Geneva*), la capitale du thé vert (the capital city of green tea*: Uji*)*. We can notice that this method enables the detection of neology. For instance, *le Bratt Pitt caché des Açores*, pure neological INE, unlikely to be found in a lexicon, has been detected.

The use of an INE lexicon enables finding patterns that do not specifically match with any rule, such as *la marque à la pomme* (*the brand with the apple*: Apple).

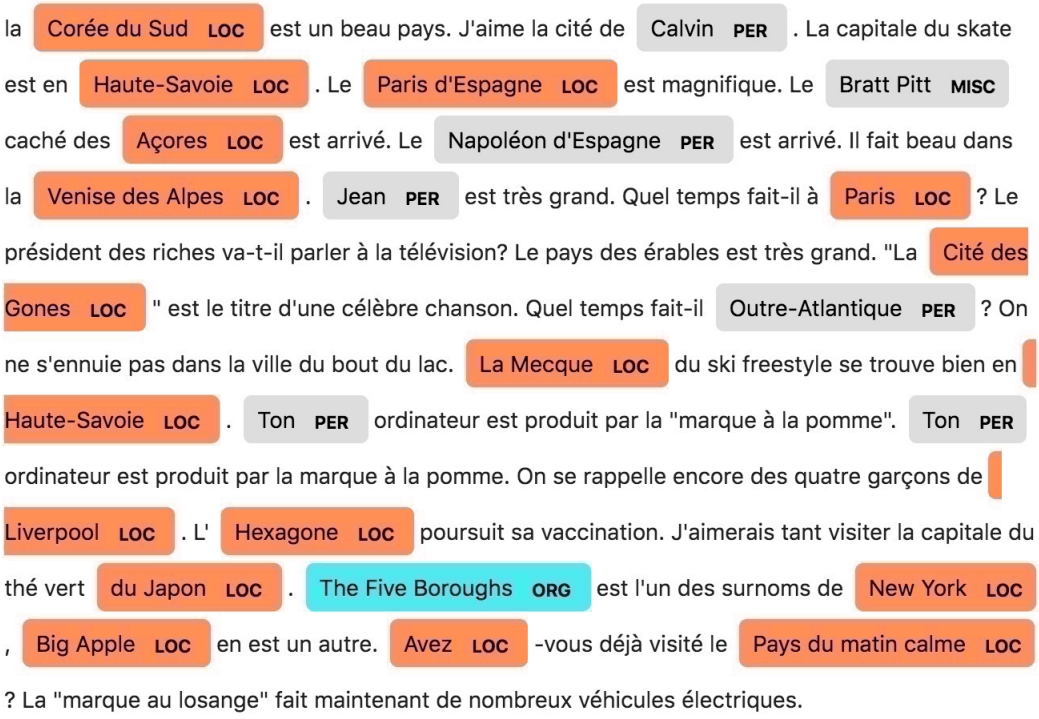

FIGURE 3. BASELINE: FRENCH NER WITH SPACY 3.0.5

Semantic tags: LOC: location, PER: person, ORG: organisation

FIGURE 4. EXAMPLE OF FRENCH NER + LEXICAL AND RULE-BASED INER

Semantic tags: LOC: location, GPE: geographical entities (similar to LOC, but detected by rules), PER: person, ORG: organisation, MISC: miscellaneous

We can also notice in this example, among the points that should be improved in our INER tool: semantic tagging mistakes (*Le Brad Pitt caché des Açores* is a person, it should not be given the "miscellaneous" tag), the handling of quotes (*la marque à la pomme* has been detected whereas *la "marque à la pomme"* has not), the detection of alternative forms of the entities listed in the lexicon (for instance, *des quatre garçons de Liverpool*, which is a contraction of *de* (of) and *les quatre garçons de Liverpool* (*the four boys from Liverpool*: the Beatles) has not been detected even if *les quatre garçons de Liverpool* was in the lexicon), and the priority that should always be given to INE (for example, *La Mecque* (Makkah), a named entity, has been detected while we would have wanted *La Mecque du ski freestyle* (*the Makkah of freestyle skiing*: la Clusaz), an INE, to be detected).

4. Discussion and further work

In order to be able to store information (such as, for instance, the referred named entity) about INE, adding to the lexicon the new INE detected by the detection rules would be useful. This would improve the lexicon and also allow a fast retrieval in cases where the lexico-syntactic and syntactico-semantic detection rules cannot be computed. For the improvement of our model for INER, our lexicon may be expanded by compiling more data from Wikipedia, as it has been done for geographic entities in French, or by doing data mining on search engine requests. Moreover, our set of implemented detection rules may be expanded and adapted to English INER.

An important question for information retrieval purposes, as discussed in Rosales-Mendez et al. (2019) and Wu et al. (2018), is *entity linking*, aiming at showing, from an INE, which named entity it refers to, and especially to solve unknown or ambiguous cases (as in Nebhi 2013). Our lexical approach takes the first step in this direction. However, in future work, we should find, among the existing methods, which one would be the best for solving ambiguous or unknown cases.

5. Conclusion

We have defined the task of INER and shown that INE can be a specific subject of study. We have proposed a first lexical and rule-based approach to retrieve INE in texts and shown its feasibility, provided that enough lexical data and the detection rules are recorded in sufficient quantity. As a result of this work, we have also compiled a dataset of INE and their equivalents and made it available to the community.

Acknowledgements

This research has been funded by the FEDER (Fonds européen de développement régional) and selected by the French-Swiss programme Interreg V.

We would like to thank Claire Wuillemin for her preliminary work in the DecRIPT project about the State-of-the-Art in NER and SER in 2020. We would also like to thank for their advice Gilles Falquet, Luka Nerima, Eric Wehrli and Jean-Philippe Goldman at the University of Geneva.

References

Abney, Steven. 1987. "The English Noun Phrase in its Sentential Aspect." PhD diss., Massachusetts Institute of Technology.

Alsharaf, H., S. Cardey, P. Greenfield, D. Limame, and I. Skouratov. 2003. "Fixedness, the complexity and fragility of the phenomenon: some solutions for natural language processing." In *Proceedings of ICL17*. Prague, Czech Republic: Matfyzpress.

Ananthanarayanan, Rema, Vijil Chenthamarakshan, Prasad M Deshpande, and Raghuram Krishnapuram. 2008. "Rule Based Synonyms for Entity Extraction from Noisy Text." In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data AND '08*, 31–38. Singapore: Association for Computing Machinery. doi:10.1145/1390749.1390756.

Bachellier, Jean-Louis. 1972. "Sur-Nom." *Le texte: de la théorie à la recherche*, no. 19: 69–92. doi :10.3406/comm.1972.1283.

Baldwin, Timothy, and Su Nam Kim. 2013. "Multiword Expressions." In *Handbook of Natural Language Processing*, Second Edition, edited by Nitin Indurkhya and Fred J. Damerau, 267–292. Boca Raton, USA: CRCPress.

Bohn, C., and Kjeti Nørvag. 2010. "Extracting Named Entities and Synonyms from Wikipedia." In *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications*, 1300–1307.

Cai, Desheng, and Gongqing Wu. 2019. "Content-aware attributed entity embedding for synonymous named entity discovery*." Neurocomputing* 329: 237–247.

Chakrabarti, K., S. Chaudhuri, T. Cheng, and Dong Xin. 2012. "A framework for robust discovery of entity synonyms." In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1384–1392, Beijing, China: Association for Computing Machinery.

Charton, Eric, Michel Gagnon, and Benoit Ozell. 2011. "Génération automatique de motifs de détection d'entités nommées en utilisant des contenus encyclopédiques (Automatic generation of named entity detection patterns using encyclopedic contents)" [in French]. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*. Articles longs, 13–24. Montpellier, France: ATALA.

Cho, Hyejin, Wonjun Choi, and Hyunju Lee. 2017. "A method for named entity normalization in biomedical articles: application to diseases and plants." *BMC bioinformatics* 18, no. 1 (1–12.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Friburger, Nathalie. 2006. "Linguistique et reconnaissance automatique des noms propres." *Meta* 51, no. 4: 637–650. doi:10.7202/014331ar.

Guenoune, Hani, Kevin Cousot, Mathieu Lafourcade, Melissa Mekaoui, and Cédric Lopez. 2020. "A Dataset for Anaphora Analysis in French Emails." In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, 165–175. Barcelona, Spain (online): Association for Computational Linguistics.

Honnibal, Matthew, and Ines Montani. 2017. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing."

Kampeera, Wannachai, and Sylviane Cardey-Greenfield. 2012. "Building a Lexically and Semantically-Rich Resource for Paraphrase Processing." In *Advances in Natural Language Processing*, edited by Hitoshi Isahara and Kyoko Kanzaki, 138–143. Springer Berlin Heidelberg.

Kauffmann, Alexis. 2013. "Structural Asymmetries in Machine Translation: The case of English-Japanese". PhD diss., Université de Genève. doi:10.13097/archive-ouverte/unige:34540.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. "Neural Architectures for Named Entity Recognition." *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, 260–270. San Diego, California: Association for Computational Linguistics.

Lin, Bill Yuchen, Dong-Ho Lee, M. Shen, Ryan Rene Moreno, X. Huang, Prashant Shiralkar, and X. Ren. 2020. "TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8503–8511. Online: Association for Computational Linguistics.

Lopez, C., Melissa Mekaoui, K. Aubry, Jean Bort, and Philippe Garnier. 2019. "Reconnaissance d'entités nommées itérative sur une structure en dépendances syntaxiques avec l'ontologie NERD." *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des connaissances, RNTI-E-35, 81–92.

Ma, Jie, Jun Liu, Y. Li, X. Hu, Yudai Pan, S. Sun, and Qika Lin. 2020. "Jointly Optimized Neural Coreference Resolution with Mutual Attention." In *Proceedings of the 13th International Conference on Web Search and Data Mining*. Houston, Texas, USA: Association for Computing Machinery.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: System Demonstrations, pp. 55-60. Baltimore, Maryland: Association for Computational Linguistics.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Benoıt Sagot, and Djamé Seddah. 2020. "Les modèles de langue contextuels CamemBERT pour le français: impact de la taille et de l'hétérogénéité des données d'entrainement (CamemBERT Contextual Language Models for French: Impact of Training Data Size and Heterogeneity)" [in French]. In *Actes de la 6e conférence conjointe Journées d'Etudes sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Etudiants Chercheurs en Informatique pour le´ Traitement Automatique des Langues (RECITAL, 22e édition)*. Volume 2: Traitement Automatique des Langues Naturelles, 54–65. Nancy, France: ATALA et AFCP.

Mitkov, Ruslan. 2014. *Anaphora resolution*. Routledge.

Mohamed, Muhidin A., and Mourad Chabane Oussalah. 2020. "A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics." *Language Resources and Evaluation* 54 : 457–485.

Nadeau, David, and Satoshi Sekine. 2007. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30: 3–26.

Nayel, Hamada A., H. L. Shashirekha, Hiroyuki Shindo, and Yuji Matsumoto. 2019. "Improving Multi-Word Entity Recognition for Biomedical Texts." *CoRRabs*/1908.05691. arXiv:1908.05691.

Nebhi, Kamel. 2013. "Named Entity Disambiguation using Freebase and Syntactic Parsing." In LD4IE@ISWC.

Nouvel, Damien, Maud Ehrmann, and Sophie Rosset. 2016. "Evaluating Named Entity Recognition." Chap. 6 in *Named Entities for Computational Linguistics*, 111–129. John Wiley & Sons, Ltd.

Ortiz Suarez, Pedro Javier, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoıt Sagot. 2020. "Establishing a New State-of-the-Art for French Named Entity Recognition" [in English]. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4631–4638. Marseille, France: European Language Resources Association.

Petit, Gérard. 2006. "Le nom de marque déposée : nom propre, nom commun et terme." *Meta* 51, no. 4: 690–705. doi:10.7202/014335ar.

Qu, Meng, Xiang Ren, and Jiawei Han. 2017. "Automatic Synonym Discovery with Knowledge Bases." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 997–1005. KDD '17. Halifax, NS, Canada: Association for Computing Machinery.

Racicot, André. 2009. "Traduire le monde: Venise du Nord et autres surnoms." *L'Actualité langagière*, vol. 6, n° 2, 23. Travaux publics et Services gouvernementaux Canada.

Rey, François-Claude, and Kauffmann Alexis. 2021. "French indirectly named entities (version 1.3) [Data set]." Zenodo. doi:10.5281/zenodo.5158253.

Rosales-Méndez, Henry, Aidan Hogan, and Barbara Poblete. 2019. "Fine-Grained Evaluation for Entity Linking." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-

IJCNLP), 718–727. Hong Kong, China: Association for Computational Linguistics.

Sales, Juliano Efson, André Freitas, Brian Davis, and Siegfried Handschuh. 2016. "A Compositional-Distributional Semantic Model for Searching Complex Entity Categories." In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 199–208. Berlin, Germany: Association for Computational Linguistics.

Schmitt, X., S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon. 2019. "A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate." *In Proceedings of the Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 338–343. doi:10.1109/SNAMS.2019.8931850.

Shang, Jingbo, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. "Learning Named Entity Tagger using Domain-Specific Dictionary." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2054–2064. Brussels, Belgium: Association for Computational Linguistics.

Shen, Jiaming, Ruiliang Lyu, Xiang Ren, Michelle Vanni, Brian Sadler, and Jiawei Han. 2019. "Mining entity synonyms with efficient neural set generation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:249–256. doi:10.1609/aaai.v33i01.3301249.

Shinyama, Yusuke, Satoshi Sekine, and Kiyoshi Sudo. 2002. "Automatic Paraphrase Acquisition from News Articles." In *Proceedings of the Second International Conference on Human Language Technology Research*, 313–318. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc.

Sjöblom, Paula. 2016. "Commercial names." Chap. V.31 in *The Oxford Handbook of Names and Naming*, edited by Carole Hough, 453–464. Oxford, UK: Oxford University Press.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. "BERT Rediscovers the Classical NLP Pipeline." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. Florence, Italy: Association for Computational Linguistics.

Treps, Marie. 2012. *La rançon de la gloire - Les surnoms de nos politiques*. Paris, France: Editions du Seuil.

Watanabe, Taiki, Akihiro Tamura, Takashi Ninomiya, Takuya Makino, and Tomoya Iwakura. 2019. "Multi-Task Learning for Chemical Named Entity Recognition with Chemical Compound Paraphrasing." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6244–6249. Hong Kong, China: Association for Computational Linguistics.

Wehrli, Eric, and Luka Nerima. 2018. "Anaphora resolution, collocations and translation." In *Multiword units in machine translation and translation technology*, edited by Johanna Monti, Violeta Seretan, Gloria Corpas Pastor, and Ruslan Mitkov, 244–256. John Benjamins.

Wehrli, Eric, Violeta Seretan, and Luka Nerima. 2010. "Sentence Analysis and Collocation Identification." In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, 28–36. Beijing, China: Coling 2010 Organizing Committee.

Weston, L., V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. 2019. "Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature." *Journal of Chemical Information and Modeling* 59, no. 9: 3692–3702. doi: 10.1021/acs.jcim.9b00470.

Wu, G., Y. He, and X. Hu. 2018. "Entity Linking: An Issue to Extract Corresponding Entity With Knowledge Base." IEEE Access 6: 6220–6231. doi:10.1109/ACCESS.2017.2787787.

Yang, Yiying, Xi Yin, Haiqin Yang, Xingjian Fei, Hao Peng, Kaijie Zhou, Kunfeng Lai, and Jianping Shen. 2021. "KGSynNet: A Novel Entity Synonyms Discovery Framework with Knowledge Graph." In *Database Systems for Advanced Applications*, edited by Christian S. Jensen, Ee-Peng Lim, De-Nian Yang, Wang-Chien Lee, Vincent S. Tseng, Vana Kalogeraki, Jen-Wei Huang, and Chih-Ya Shen, 174–190. Cham: Springer International Publishing.

Zhang, Ruoyu, Wenpeng Lu, Shoujin Wang, Xueping Peng, Rui Yu, and Yuan Gao. 2021. "Chinese clinical named entity recognition based on stacked neural network." *Concurrency and Computation: Practice and Experience* : 33:e5775. doi:10.1002/cpe.5775.