Named Entity Resources - Overview and Outlook

Maud Ehrmann, Damien Nouvel, Sophie Rosset

Named Entity Resources - Overview and Outlook

Maud Ehrmann[1], Damien Nouvel[2], Sophie Rosset[3]
[1] Digital Humanities Laboratory, EPFL, Switzerland
[2] ERTIM, Inalco, Paris, France
[3] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
maud.ehrmann@epfl.ch, damien.nouvel@inalco.fr, sophie.rosset@limsi.fr

Abstract

Recognition of real-world entities is crucial for most NLP applications. Since its introduction some twenty years ago, named entity processing has undergone a significant evolution with, among others, the definition of new tasks (e.g. entity linking) and the emergence of new types of data (e.g. speech transcriptions, micro-blogging). These pose certainly new challenges which affect not only methods and algorithms but especially linguistic resources. Where do we stand with respect to named entity resources? This paper aims at providing a systematic overview of named entity resources, accounting for qualities such as multilingualism, dynamicity and interoperability, and to identify shortfalls in order to guide future developments.

Keywords: named entity, linguistic resources, NE typologies, annotated corpora, evaluation, linked data

1. Introduction

Recognition of real-world entities is crucial for most, if not all, text mining applications. Indeed, referential units such as name of persons, places and organisations underlie the semantics of texts and guide their interpretation. Acknowledged some twenty years ago, named entity (NE) mining is an operation of ever increasing importance for many NLP applications which has undergone major evolutions since then.

Named entity processing is representative of the evolution of information extraction from a document to a semantic-centric view point (Rao et al., 2013). As first introduced during the 6^{th} Message Understanding Conference (Grishman and Sundheim, 1995), it corresponds to the recognition and classification of entities of interest in texts, generally of type *Person*, *Organisation* and *Location*. During subsequent evaluation campaigns and research programs, the task quickly broadened and became more complex, with the extension and refinement of typologies (Sekine et al., 2002; Galibert et al., 2011), the diversification of languages taken into account (Tjong Kim Sang and De Meulder, 2003; Santos et al., 2006; Magnini et al., 2008a; Galliano et al., 2009; Benikova et al., 2014), and the expansion of the linguistic scope with, along proper names, the consideration of nominal phrases and pronouns as candidate lexical units (Doddington et al., 2004). Later on, as recognition and classification were reaching satisfying performances (at least for well-covered languages and main stream texts such as news articles), attention focused on finer-grained processing, e.g. metonymy recognition (Markert and Nissim, 2009), and on the next logical step, namely disambiguation.

Entity resolution has first been defined as a clustering problem, where different mentions in texts referring to the same entity must be grouped together (Mann and Yarowsky, 2003; Artiles et al., 2008). Next, with the advent of knowledge bases (KB) containing plenty of entities along with detailed information (Hovy et al., 2013), entity disambiguation switched from clustering mentions to aligning mentions to unique identifiers in a KB. Thereupon, the task of entity linking gained strong impetus (Ji and Grishman, 2011; Shen et al., 2015) and is now at the core of many knowledge extraction tools for the Semantic Web (Gangemi, 2013).

In addition to this task-related vertical evolution, NE processing also branched out into several directions. Besides the general domain of well-written news wire data, work is carried out on specific domains, particularly bio-medical (Kim et al., 2003), and on more noisy input such as speech transcriptions (Galibert et al., 2014) and tweets (Ritter et al., 2011). Likewise, NE processing is called on to contribute in other research areas such as Digital Humanities, with OCRed documents (Rosset et al., 2012a) and languages or documents of earlier stages (Rodriquez et al., 2012; Brando et al., 2015).

NE processing encompasses therefore different tasks which can be performed within a large variety of contexts. Accordingly, different methods and algorithms are used which, regardless of their nature, all require resources encoding linguistic knowledge about these units. Over the last decades, many named entity resources have been built, addressing different needs, for different languages and for various input data. In this regard, the above mentioned development of NE-related tasks (e.g. entity linking) and the emergence of new input data (e.g. social media, multimedia, ancient texts) pose certainly new challenges, while the availability of rich knowledge bases (e.g. DBpedia) represent new opportunities.

What resources are available for which task and in which context? Where can they be found and in which format? How to assess their quality? In the context of growing interest for language resource publication and sharing, this paper presents an overview of linguistic resources for named entities. More precisely, our objectives are to:

- provide an account of NE resources as complete and up-to-date as possible (community general knowledge perspective), and
- identify NE resources shortfalls in order to guide future developments (road-map perspective).

The remainder of this paper is organized as follows: after the discussion of related work (section 2.), we introduce major NE resource types in relation to NE tasks. Next, we describe our assessment approach (section 4.), considering both assessment criteria (section 4.1.) and methodological aspects (section 4.2.), and propose a systematic overview of NE resources (section 5.). Finally, we discuss needs and priorities for named entity resources (section 6.) and conclude (section 7.).

2. Related work

Language resources (LR) have long been acknowledged as a cornerstone of NLP processes and the number of published resources is constantly growing. Although extremely valuable, the resulting set of language data is however difficult to handle and several initiatives have emerged to facilitate the discovery, the search and the documentation of LRs. Initiated in 2010, the LRE map intends to enhance the availability of information about resources (Calzolari et al., 2012) while large-scale projects such as CLARIN and META-SHARE aim at indexing LR repositories into far-reaching networks. Despite these efforts, the discovery of resources remains a difficult process for metadata vocabularies differ from each other. Taking advantage of RDF, a recent work by (McCrae et al., 2015) attempts to harmonize heterogeneous descriptions of language resources.

Beyond discoverability, the Open Linguistics Working Group[1] concentrates on linguistic data interoperability by considering the Linked Data paradigm as a way to reconcile data and metadata descriptions (Chiarcos et al., 2012; Heath and Bizer, 2011). This community-based effort has initiated the creation of the Linguistic Linked Open Data Cloud and recently launched the metadata repository LingHub[2].

While these enterprises are interested in linguistic resources as a whole, the present work focuses instead on resources related to named entities. The objective is not only to describe what exists, but especially to explain what serves what and to assess the extent to which today's NE processing requirements are met in terms of resources. The study presented by (Nadeau and Sekine, 2007) discusses named entity recognition and classification at large. More recently, a deliverable from the QTLeap project surveys NE disambiguation tools and datasets (Agirre et al., 2015). Besides these two studies, named entity resources descriptions are scattered over articles describing single tools and/or resources. To the best of our knowledge, there is no complete description and assessment of NE resources.

3. Named entity tasks and resources

Named entity family of tasks cover logical steps of text processing in increasing level of complexity and are defined as follows:

- **recognition**: detecting named entities, i.e. elements in texts which act as rigid designator for a referent[3],
- **classification**: categorising named entities according to a set of pre-defined semantic categories,
- **disambiguation/linking**: linking named entity mentions to a unique reference, and
- **relation extraction**: discovering relations between named entities.

There exists two main usages of named entities depending on whether the application focuses on referential entities (e.g. indexing and knowledge integration), or whether it focuses on mentions (e.g. knowledge base population, anonymisation and information extraction at large) (Fort et al., 2009). Usual NE task combinations reflect these usages: recognizing and classifying are part of *named entity recognition* (suitable when focusing on mentions) and recognizing and disambiguating are part of *entity linking* (suitable when focusing on referents).

For most NE systems and algorithms, resources are crucial for the achievement of these tasks. Three main types of resources may be distinguished, each playing a specific role. **Typologies** are used to define a semantic framework for the entities under consideration and are required for classification. They can be multi-purpose or domain-specific and of various degree of hierarchization. Next, **lexicons and knowledge bases** are responsible for providing information about named entities and are used for recognition, classification and disambiguation. This information is either of lexical nature, relating to the units making up named entities, or encyclopedic, concerning their referents. Finally, **annotated corpora** are used to illustrate an objective, and may be used as a learning base or as a point of reference for evaluation purposes.

4. Resource assessment approach

We consider under which aspects named entity resources should be considered and describe our review methodology.

4.1. Assessment criteria

We consider the following **criteria**, which we believe can assess essential *qualities* of named entity resources:

- **Language**, in order to assess *multilingualism*. How complete is the language coverage of NE resources? The existence and availability of NE resources in multiple languages is a core concern to carry research on NE and develop NE-based applications.
- **Domain**, in order to assess *applicability*. Which spheres of activity do NE resources cover? Typologies and type of lexical units can vary significantly from one domain to another.
- **Type of text**, in order to assess *robustness*. Are NE resources available for all kind of texts? Beyond the distinction regarding modality with written vs. spoken data, different types of text such as microblogging vs. news articles vs. broadcast conversations should be taken into account.

[1] http://linguistics.okfn.org

[2] http://linghub.org

[3] The question of what should be considered as named entity is not the point in question here.

- **Update capabilities**, in order to assess *dynamicity*. How likely do NE resources become obsolete? Named entities are constantly changing and resources should be updated accordingly.

- **Format and vocabulary**, in order to assess *interoperability*. Are NE resources interoperable and at which level? Combined use as well as reuse of NE resources depends on the ability to syntactically process and semantically interpret information in a seamless way (syntactic and semantic interoperability).

- **License**, in order to assess *transparency* and *openness*. When specified, under which license NE resources are available?

- **Quality**, in order to assess *efficiency* and *accuracy*. How reliable is the information provided by named entity resources? Intrinsic quality of resources depends on various factors such as the nature of data used to build the resource (texts, web pages, Wikipedia), the building approach (manual, semi-supervised, unsupervised), the potential use of pre-processing tools (e.g. Optical Character Recognition or Automatic Speech Recognition systems) or the quality of annotation campaigns (Fort, 2012).

4.2. Methodology

Compiling information about named entity resources is neither an easy nor a neutral process: complete coverage cannot be guaranteed and there are certainly biases. The adopted methodology intends to minimise such pitfalls.

As NE processing originates from evaluation campaigns, the first option to find information is to screen such events for the past and present times. Another option is to search for named entity-related data in linguistic resource catalogues, such as LRE map, LDC and ELDA. Still, evaluation campaign and distribution agency websites do not list all information; it is often necessary to review research articles on the topic and to examine research institutions and researchers' web sites to complement the findings.

For people familiar with NE processing, this process is not as tedious as it seems. It might however contain biases, mainly affecting the survey coverage. The first one corresponds to the fact that we might cover only well-known and well-visible sources of information, with the consequence of missing isolated ones. The second corresponds to the fact that we are surely influenced by our backgrounds (mostly general domain), with the consequence of missing resources for other domains. If both biaises can be mitigated by contacting researchers that could help us to further complete our findings, we must not fool ourselves: gathering a *comprehensive* and *precise* picture of NE resources is almost impossible. However, it seems reasonable and certainly to be of benefit to aim at an overall picture.

In practical terms, we completed our inventory process according to the three above mentioned exploration strategies. We looked exclusively at the types of resources identified in section 3. (typologies, corpora and lexicons/KBs) and ignored NER tools and web services[4]. With respect to the screening of resource catalogs, we extensively used the recently set up LingHub portal which gather information about language resources in RDF format from CLARIN, LRE Map, META-SHARE and Datahub[5]. The SPARQL endpoint of the portal proved to be very helpful and saved us time by serving appropriate pointers and information on resources. We complemented these searches by looking at evaluation campaigns, at the LDC and ELDA catalogs and at individual publications on resources. It is worth mentioning that many publications mention the creation of a NE-resource without, unfortunately, releasing it.

5. Resources for Named Entities

Based on our inventory results (state February 2016), this section surveys NE resources according to the assessment criteria defined in section 4.1..

We were able to inventory many corpus (121 items), quite a few lexicons/KBs (29) and a bit less typologies (23). These proportions reflect the role (but not the importance) and the usage of each type of resource. We compiled three tables offering a synthetic view of each resource type, comprising several descriptive properties. These include, for corpora: the program which financed the resource (if any), the resource modality (written, spoken, both), the textual genre (web, news, social media, etc.), the domain (generic, specific), the language, the type of tagging (part-of-speech, semantic roles, named entities, etc.), the annotation modality (manual, automatic, semi-automatic), the typology used, the size, format, license, and pointers to websites and references. In the case of lexicons/KBs, we additionally consider whether they are regularly updated or not, the source they where compiled from and the type of their content (entity names and/or trigger words). As for typologies, the number of categories, the existence of sub-types and of nested entities and the type of lexical unit to consider as NE are also taken into account. It is important to note that not all information is not always available for a resource. The three tables, which might be further complemented in the future, are available for consultation on-line[6]. They form the basis of the following observations.

5.1. Typologies

In the context of named entities, a typology corresponds to a formalized and structured description of the semantic categories to consider (the objects of the world which are of interest), along with a definition of their scope (their realization in texts). There exists different typologies, most of them defined and published – usually with a few years of delay– as part of evaluation campaigns, with no tradition of releasing typologies as such outside this context.

Typologies differ in their definition of semantic categories (scope), in their degree of extensiveness (more or less cate-

[4]Those are more the focus of frameworks such as NERD (`http://nerd.eurecom.fr/`) or GERBIL (`http://aksw.org/Projects/GERBIL.html`)

[5]`http://www.clarin.eu` `http://metashare.elda.org` `http://datahub.io`

[6]`http://damien.nouvels.net/resourcesen/{typologies.html,lexicon.html,corpora.html}`

gories) and of granularity and hierarchisation (more or less structuring). It is indeed possible to distinguish between "simple" typologies comprising only a few categories (say up to 5) and complex ones with numerous classes. Typologies defined in the context of MUC, CoNLL and EVALITA (Grishman and Sundheim, 1995; Tjong Kim Sang and De Meulder, 2003; Magnini et al., 2008b) fall within the first group, while the typology of Sekine (Sekine et al., 2002) and the ones established for the HAREM, ETAPE, ESTER and GERMEVAL (Santos et al., 2006; Galibert et al., 2014; Galliano et al., 2009; Benikova et al., 2014) campaigns belong to the second. Besides, some typologies offer finer-grained classification than others: while MUC, ACE and CoNLL do not distinguish subtypes, ESTER and especially QUAERO (Rosset and Grouin, 2011) define significantly more specific categories. The latter considers what is more the components (function, title, etc.) which make up named entities. Finally, the definition of what do semantic classes cover can greatly diverge from one typology to another with, first, different definitions of entity spans (e.g. inclusion or not of person titles) and, second, different appreciations of what should be considered as a named entity (e.g. proper names, definite descriptions, pronouns).
In the following we analyse typologies according to our assessment criteria. All do not apply for this type of resource.

Language Typologies can, to a large extent, be considered as language-agnostic since the definition of objects of interest do not change depending on the language. If one could hypothesize that different cultural backgrounds will end up with different semantic categories, this can hardly be verified based on existing typologies which, for the most part, come from the western world. In case of typologies, language relates as well to how semantic categories are lexicalized; in this regard, English is well-resourced while some European languages for which a campaign has been held have a unique typology (de, it, pt) . In the case of "simple" NE typologies, lexicalization language does not really affect their usability.

Domain The domain covered by the typology, in turn, has a huge impact on their applicability, as typologies from different domains cannot be interchanged. It appears that most of them are for the general and bio-medical domains; however, they are not all published and private text mining applications have probably specialized typologies.

Update capabilities Regarding updating capacities, traditional NE typologies are somehow dynamic in the sense that, from one campaign to another, they inherit from each other and are gradually adjusted. It is worth noticing that old typologies such as MUC or CoNNL are still used today (cf. 'used typology' property in the table on corpora). Recently, a new type of NE classification has emerged from the development of the on-line collaborative encyclopedia Wikipedia and its linked data counter-parts DBpedia and Wikidata. Indeed the underlying ontologies, with numerous classes, serve more and more as a basis for the selection of entity types, particularly in the context of entity linking.

Format Typologies are often published as part of annotation guidelines defined during annotation campaigns. As such, it is a resource that is mostly meant for humans, in the context of annotation and evaluation purposes. Recent work has been conducted to establish XML standards to formalize typologies, mainly in the context of corpus annotation (e.g. Folia[7]). The recent use of ontologies brings also into play the Web Ontology Language (OWL).

Quality Assessing the quality of a typology depends on the targeted application and on the domain. In this context, typologies can be evaluated according to whether or not they meet the requirements of an IE or text mining application. If simple typologies are the most wide-spread, one is however entitled to wonder whether it is because they answer all needs, or because NE processing tools cannot recognize more complex objects. Regarding this criteria, it should moreover be noted that the methodology used to build typologies (top-down, bottom-up or mixed) is almost never reported, with the exception of (Sekine et al., 2002).

5.2. Corpora

Annotated corpora correspond to sets of documents enriched by named entity tagging according to a given typology. They are essential in the context of developing and evaluating systems for NE recognition, classification and disambiguation. This type of resource is the one for which we found the most instances, with 121 inventoried resources.
In order to get a better overall picture, Figure 1 presents a visualization of the NE-annotated corpora that we inventoried for the general domain. Information is rendered using circles which represent the amount of all available data given (a) a language (showed on the y-axis), (b) a resource modality (written, spoken, mixed or other, showed with the inner color) and (c) an annotation procedure (manual, automatic, semi-automatic or unknown, showed with the circumference color). This means that, for example, *all* English corpora composed of written data and annotated according to an unknown annotation procedure are represented with the red circle with a grey pourtour (the second circle on the 'en' line). The size of the circle corresponds to the size of corpora, normalized on the total size of all available corpora across languages. As with typologies, we confront NE-annotated corpora with the criteria. For a detailed presentation, we refer the reader to the on-line table[8].

Language Language coverage of NE-annotated corpora has a huge impact on the development of NE processing tools. Indeed, although new unsupervised approaches based on deep learning are being developed (Dos Santos and Guimaraes, 2015), annotated corpora are widely used to train and evaluate NER and EL systems. Our inventory features corpora in 17 languages. The most resourced language is, not surprisingly and whatever the modality and annotation procedure, English; the less-resourced one is Basque (cf. Figure 1). Besides, western and eastern European languages are quite well covered while African, Semitic and Asian languages have less annotated material. Thus, if the degree of *multilingualism* of NE-annotated corpora is not catastrophic, it still remains an issue as many languages are not covered at all.

[7] http://proycon.github.io/folia
[8] See footnote 6.

Domain The vast majority of corpora are of the general domain and there exist a fair proportion of data for the biomedical domain. Beyond that, some corpus focus on specific domains such as 'crime' (*Reuters 128*) or 'sciences' (*RSS 500*) but this remains rather an exception. In our inventory, English has the more diverse domains. It should therefore be concluded that the degree of *applicability* of available NE-annotated corpora is rather low.

Type of text The great majority of corpora contain data of written modality (78.6%[9]), a minority of spoken modality (15.4%) and a few of the two (6%) with only 6 languages (ar, en, fr, it, pt and zh[10]). Regarding the textual genre, a bit less than half of the corpora (45.4%) comes from the news domain (including broadcast conversations and news, news papers and news wire). Web data is the next most represented (15%), followed by scientific material (12%, composed of abstracts, articles, recorded seminars) and Wikipedia-derived data (12%). The less represented types of text are Social Media (tweets, SMS, short messages) with only 8.3%, and cultural heritage texts, with only very few corpora, e.g. (Rosset et al., 2012b).

Update capabilities In the context of corpora, update capabilities refer to the fact that material can be re-annotated with a new or revised typology and/or a different annotation procedure. This remain quite rare and corpora which were built a decade ago are still in use (e.g. MUC and CoNLL data). However, the advent of the entity linking task has recently led to the re-annotation of old corpora, adding entity references on top of entity mentions (Hoffart et al., 2011). During the QUAERO annotation campaign, the French ESTER corpus got as well re-annotated with the QUAERO typology. Besides, these updates do not account for the evolution of language itself; new data is needed to gather new entities, e.g. new proper names, and also new ways to refer to entities in texts (e.g. using the `@` and `#` characters from Twitter is becoming popular).

Format and vocabulary Format affects the possibility of sharing and re-using corpora, as well as NER systems' performances, as demonstrated in (Ratinov and Roth, 2009). Here we should notice that the information about corpora format is not always given (only 86 corpus out of 121 has it) and is most of the time difficult to find. Corpora of the 1990's were released in SGML, then XML spread, as well as the CoNNL BIO (Begin, Inside, Outside) - also known as IOB - annotation representation format. XML, tabular plain text and BIO are well represented in our inventory, even though some corpora are released in the form of SQL databases with dedicated APIs. The advent of the linked data publishing paradigm and the definition of appropriate vocabularies to represent linguistic information, e.g. NIF (Hellmann et al., 2013), has initiated the released of corpora in RDF format. They however remain a small minority with, as of today, only 9 NIF NE-annotated corpus (ca. 10%).

[9]Percentages are calculated in relation to the absolute number of corpora.

[10]Arabic and Chinese have to be considered as multiple of languages or dialects.

Furthermore, the vocabulary used to annotated texts, i.e. the typologies, greatly affects the usability of annotated corpora: even if several corpora exist for the same language, they often cannot be used together for they were annotated according to different typologies. The NERD framework (Rizzo and Troncy, 2012) intends to circumvent this drawback by defining a general ontology to which several existing NE typologies are mapped. Similarly as for format, the information of the typology used during the annotation is not always present (77 out of 121). In this context, statistics cannot be fully representative but can give an idea about the ratio between different used typologies: 32% of corpora are annotated with CoNNL, 14.6% with 'in-house' typologies with numerous categories (mentioned as 'extended' in the table), 12% with ACE, 9% with the three basic categories (Person, Location, Organisation), 6,6% with MUC and 5% with QUAERO. This confirms the wide adoption of CoNLL, which BIO format and Miscellaneous category seem to be well appreciated.
Finally, we should as well consider the task for which corpora have been built. This information is available for 110 corpora. The vast majority were annotated for the purpose of NERC (87%), some for entity linking (16%) and entity relation detection (15%)[11].

License The type of license has obvious implications with regards to the usage of annotated corpora. On this point we were able to retrieve information for 96 cases out of 121. License description varies significantly from one catalogue to another, thereby we propose to simply distinguish between free of charge corpora vs. those with a license fee. The first case represent 54% of the items, the second 43%.

Quality The quality of (NE) annotated corpora is difficult to assess for it depends on various parameters for which information is not always available. First, the annotation guidelines used for the annotation has a significant impact on the quality and the coherence of the annotation. Second, the way the annotation campaign is carried out, that is to say how annotators are trained, how adjudication of conflicting cases is done and how inter-annotator agreement is measured, affects also the resulting data. Finally, the quality of annotated corpora can be affected by the potential use of automatic pre-processing and/or annotation tools. In order to allow an informed use of data all this information should be documented and released along with corpora; however, it is not always the case. In our inventory, the information of the annotation procedure is available for 63 items (rest is unknown), which distribute as follows: 75% manual, 18% automatic and 12% semi-automatic. Still, given the fact that they represent only half of the inventoried corpora, we should be careful of not drawing abusive conclusions.

5.3. Lexicons and Knowledge Bases

Lexicons and knowledge bases provide information relating to entities which may be used by automatic systems for the purposes of recognition, classification and disambiguation. This type of resource has evolved significantly

[11]The sum of percentages is not 100 due to annotation type overlappings.

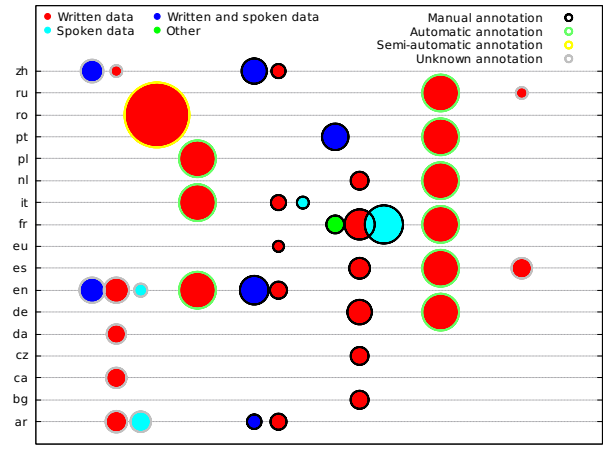

Figure 1: Visualization of inventoried NE-annotated corpora for the general domain.

in the last years, as a result of the increased complexity of NE-related tasks and of technological advancements and progress made in terms of knowledge representation with the emergence of the Semantic Web. Information relating to named entities was initially of lexical nature and stored as simple word lists for each possible category. Those lists were named "gazetteers", a term initially devoted to toponyms that was afterwards extended for any named entity type that is to be collected in texts. These lexicons encode two type of information: entity names, to use in look-up procedures, and trigger words, to use as features to guess names in texts. Later on, the consideration of ambiguity (i.e. an entity name referring to different entities), of homonymy (i.e. different names referring to the same entity) and the evolution of typologies (more structured as hierarchies or ontologies) led to the structuration of NE-related information into knowledge bases. Since the advent of Wikipedia as a primary source for NLP resources, semantic KBs were particularly boosted.

We were able to inventory 34 lexicons and KBs, which are as well available for consultation on-line[12].

Language Although entity names are often not translated from one language to another, they vary according to scripts and have multiple variants, within and across languages. Lexicons of entity names should therefore exist in different scripts and include multilingual name variants. Regarding trigger words (e.g. *Mister*, *Pope*, *the former Minister*, *Ltd.*), they differ from one language to another. The 34 collected lexicons and KBs gather data for 10 distinct languages (ar, bg, de, du, en, es, fr, hu, it and pl) to which must be added *multilingual* resources with up to 270 languages. Within this set, 8 resources are for English and 5 with multiple languages. The majority of other languages have one or two resources, with the exception of Polish with 4 items. Similarly as for annotated corpora, European languages are well represented. Regarding their content, almost all collected resources contains entity names while only 5 contains triggers words.

[12]See footnote 6

Domain With the exception of the Geonames resource (dedicated to place names) and of bio-medical lexicons – a quite well covered domain –, all resources are for the general domain.

Type of text At first sight one can think that lexicons and KBs can be used interchangeably for almost all type of texts. This should be qualified, however: social media material such as tweets are plenty of truncated and/or abbreviated units, and real-life texts contains typos and unexpected variants. On this point, it is clear that these issues are poorly covered by existing resources; only the JRC-Names resource (Steinberger et al., 2011; Ehrmann et al., 2016) offer units gathered from real-life data.

Update capabilities Composed mainly of proper names, named entities evolve endlessly, as new people, places, products, companies, etc. appear and disappear. It is thus crucial that lexicons and KBs used for their recognition and disambiguation are regularly updated. Given the fact that more an more resources are now derived from Wikipedia or can be contributed to (e.g. Geonames), updating capabilities become more common than in the past. Out of 34 resources, 10 are "frozen" while 21 are (or can be) regularly updated.

Format and License As for the format, lexicons and KBs are as follows: on 34 resources, 23 include this information, with 12 resources in RDF, 5 in LMF, 3 as plain text and 2 in SQL. When mentioned, license is most of the time an open one.

Quality In the context of lexicons and KBs, quality can only be assessed through extrinsic evaluation, i.e. while using the resource for NE processing. Indeed, neither a large resource nor a manually built one guarantee good performances. On this point, it should be stressed that more and more resources (DBpedia, Yago, BabelNet) are built from Wikipedia data. Wikipedia is good at providing up-to-date and multilingual data, but is somehow bound to VIP entities[13].

6. From a map to a roadmap

We first recapitulate the situation, by quickly summarizing the main observations on NE resources according to the assessment criteria, and then discuss the needs and priorities for future developments.

Language Typologies are little concerned with multilingualism; NE-annotated corpora show a modest degree of multilingualism and a similar situation holds for lexicons and KBs which, despite a few highly multilingual resources, are not available for all languages.

Domain Typologies are mainly available for the general and bio-medical domain; corpora follow the same line, as well as lexicons and KBs. NE resources applicability is therefore rather low.

Type of text Typologies are not concerned with this criteria. Regarding corpora and lexicons/KBs, the same observation can be made: majority of written data and of news

[13]In the TAC-KBP 2010 query set, 57% of queries (entities) were missing from the KB (Rao et al., 2013).

genre; deficit of spoken, social media and cultural heritage data

Update capabilities Typologies are updated thanks to inheritance phenomena and to ontologies derived from online collaborative resources. Corpora are slightly updated and new ones are built. Lexicons and KBs are surprisingly well updated. This means that the dynamicity of NE resources is quite acceptable

Format and vocabulary Corpora have diverse formats and different vocabularies, which hinder from using them conjointly. Lexicon and KBs use LMF and RDF.

License This criteria does not apply for typologies. Only a bit more than half of the corpora are freely available. The situation is better for lexicons and KBs.

Quality Typologies quality can be assessed via their usage. Building approaches of corpora are generally little documented. Lexicons and KBs allow to perform more complex task than before (EL), but no objective evaluation of their quality exists (e.g. by comparing performances obtained with different resources on the same data and with the same tool).

In the light of the above, it appears that NE resources are quite well developed, benefiting from two decades of work. However, they do not meet all needs in terms of, first, tool efficiency and, second, text mining applications. The former could be improved by an effort of *harmonization* for both format and vocabularies. These aspects relate mainly to decisions which can be taken at early stages of projects. The latter could be enhanced by further *developments* of resources for more languages, different types of texts and different domains. These aspects appertain, in turn, to mid or long term development. In this regard, where to guide future developments should estimate the combinatorics of the multiple axis considered (language, domain, type of text) as it is certainly not feasible to develop everything. Finally, this raises several issues such as: if we want to be able to work in all languages, all domains, with all language-stages, what is the most important and should be given priority? Should the research community focus only on feasibility studies/cases and let others develop what is needed in specific contexts or, on the contrary, try to make advances in all directions? As future work, we intend to further discuss these points, based on an even more complete inventory.

7. Conclusion

We have presented an overview of existing NE resources, accounting for a series of description and assessment criteria. The resources we inventoried helped us to identify gaps in the NE-resource landscape and to suggest directions for future developments. We welcome any feedback that could help complement this review.

Agirre, E., Alegria, I., Artetxe, M., and et al., N. A. (2015). Report on the State of the Art of Named Entity and Word Sense Disambiguation. Deliverable D5.1 Version 4.0. Technical report, QTLeap Project.

Artiles, J., Sekine, S., and Gonzalo, J. (2008). Web people search: results of the first evaluation and the plan for the second. In *Proc. of the 17th international conference on World Wide Web*, pages 1071–1072. ACM.

Benikova, D., Biemann, C., and Reznicek, M. (2014). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proc. of the* 9^{th} *International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Brando, C., Frontini, F., and Ganascia, J. (2015). Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets. In *New Trends in Databases and Information Systems*, pages 505–514.

Calzolari, N., Gratta, R. D., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE map. Harmonising Community Descriptions of Resources. In *LREC*, pages 1084–1089.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). The Open Linguistics Working Group of the Open Knowledge Foundation. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics*, pages 153–160. Springer Berlin Heidelberg.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The ACE program, tasks, data, and evaluation. In *Proc. of the* 4^{th} *International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal.

Dos Santos, C. and Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. In *Proc. of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July.

Ehrmann, M., Jacquet, G., and Steinberger, R. (2016). JRC-Names: Multilingual Entity Name variants and titles as Linked Data. *Semantic Web Journal*.

Fort, K., Ehrmann, M., and Nazarenko, A. (2009). Towards a Methodology for Named Entities Annotation. In *Proc. of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 142–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Ph.D. thesis, Université Paris 13.

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., and Quintard, L. (2011). Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In *Proc. of 5th International Joint Conference on Natural Language Processing*, November.

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

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586.

Gangemi, A. (2013). A comparison of knowledge extraction tools for the Semantic Web. In Cimiano, P., Corcho, O., Presutti, V., Hollink, L., and Rudolph, S., editors, *The Semantic Web: Semantics and Big Data*. Springer.

Grishman, R. and Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proc. of a Conference Held in Columbia, Maryland*.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *Proc. of the* 12^{th} *International Semantic Web Conference (ISWC13)*, Sydney, Australia, October.

Hoffart, J., Yosef, M. A., Bordino, I., and et al., H. F. (2011). Robust disambiguation of named entities in text. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1148–1158. Association for Computational Linguistics.

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

Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., and et al., A. M. (2008a). Evaluation of natural language tools for italian: Evalita 2007. In *Proc. of the* 6^{th} *International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., Mazzei, A., Lombardo, V., Bertagna, F., Calzolari, N., Toral, A., and et al., V. B. (2008b). Evaluation of Natural Language Tools for Italian: EVALITA 2007. In *Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proc. of the* 7^{th} *Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CoNLL'03, pages 33–40.

Markert, K. and Nissim, M. (2009). Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

McCrae, J., Cimiano, P., Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015). Reconciling Heterogeneous Descriptions of Language Resources. *ACL-IJCNLP 2015*.

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

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

Ratinov, L. and Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ritter, A., Clark, S., Etzioni, M., and Etzioni, O. (2011). Named Entity Recognition in Tweets: An Experimental Study. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Rizzo, G. and Troncy, R. (2012). NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Rodriquez, K., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *KONVENS*, pages 410–414.

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

Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., and Zweigenbaum, P. (2012a). Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proc. of the Sixth Linguistic Annotation Workshop*, pages 40–48, Jeju, Republic of Korea, July. Association for Computational Linguistics.

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

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proc. of the* 5^{th} *International Conference on Language Resources and Evaluation (LREC'06)*, pages 1640–1643, Genoa.

Sekine, S., Sudo, K., and Nobata, C. (2002). Extended Named Entity Hierarchy. In *Proc. of the* 3^{d} *International Conference on Language Resources and Evaluation, LREC*, Las Palmas, Canary Islands, Spain.

Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460.

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

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proc. of the* 7^{th} *Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CoNLL'03, pages 142–147.