Named Entity Recognition: Fallacies, Challenges and Opportunities

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado,
Jorge Morato, Juan Miguel Gómez-Berbís

University Carlos III of Madrid
Department of Computer Science
Avda. Universidad, 30
28911 Leganés, Madrid, Spain
mmarrero@inf.uc3m.es jurbano@inf.uc3m.es sscuadra@bib.uc3m.es
jmorato@inf.uc3m.es juanmiguel.gomez@uc3m.es

Abstract. Named Entity Recognition serves as the basis for many other areas in Information Management. However, it is unclear what the meaning of Named Entity is, and yet there is a general belief that Named Entity Recognition is a solved task. In this paper we analyze the evolution of the field from a theoretical and practical point of view. We argue that the task is actually far from solved and show the consequences for the development and evaluation of tools. We discuss topics for further research with the goal of bringing the task back to the research scenario.

Keywords. Named Entity; Named Entity Recognition; Information Extraction; Evaluation methodology, NER tools

1 Introduction

Named Entity Recognition (NER) is a task in Information Extraction consisting in identifying and classifying just some types of information elements, called Named Entities (NE). As such it, serves as the basis for many other crucial areas in Information Management, such as Semantic Annotation, Question Answering, Ontology Population and Opinion Mining. The term Named Entity was first used at the 6th Message Understanding Conference (MUC)[1], where it was clear the importance of the semantic identification of people, organizations and localizations, as well as numerical expressions such as time and quantities. Most of the NER tools nowadays keep considering these types of NE originated in MUC, though with important variations.

But it is unclear what the meaning of NE is. This question has not been analyzed with detail yet, although it has manifested in other works: "the concept of NE appeared in an environment of NLP applications and is far from being linguistically clear and settled" [2]. This is oddly contrasted with the common statement that NER is a solved task with success ratios over 95% [3]. In this paper we argue that NER is in fact not a solved problem, and show how the lack of agreement around the concept of NE has important implications for NER tools and, especially, for their evaluation. Current evaluation forums related to NER do not solve this problem basically because they deal with very different tasks. The true evolution of NER techniques requires for us to reconsider whether NER is really a solved problem or not, and design evaluation procedures suitable for the current needs and to reliably assess where the NER state of the art is at.

The remainder of the paper is organized as follows. The next Section examines the application of NER, followed by a description of the evolution of the field. In Section 4 we discuss and analyze the discrepancies among several definitions for NE. Next, Sections 5 and 6 show the implications of the lack of a proper definition for both the evaluation forums and the tools. In Section 7 we reopen the question of NER being really a solved problem and conclude it is not. In Section 8 we discuss the current evaluation forums and, showing they do not tackle the same task, Section 9 presents the challenges and opportunities for continuing and refocusing research in NER. Finally, Section 10 concludes with final remarks.

2 Why do we need Named Entity Recognition?

According to a market survey performed by IDC [4], between 2009 and 2020 the amount of digital information will grow by a factor of 44, but the staffing and investment to manage it will grow by a factor of just 1.4. Dealing with the mismatch of these rates is a challenge, and one of the proposals to alleviate the problem is the development of tools for the search and discovery of information, which includes finding ways to add structure to unstructured data. Named Entity Recognition, which purports to identify semantics of interest in unstructured texts, is exactly one of the areas with this goal, serving as the basis for many other crucial areas to manage information, such as semantic annotation, question answering, ontology population and opinion mining:

- Semantic annotations go beyond familiar textual annotations about the contents of the documents to the formal identification of concepts and their relations. For example, a semantic annotation might relate the term *Paris* to an ontology identifying it as an instance of the abstract concept *City*, and linking it to the instance *France* of the abstract concept *Country*, thus avoiding any ambiguity as to which *Paris* the text refers to. These annotations, intended primarily for their use by machines, bring two main benefits for these systems: enhanced information retrieval and improved interoperability [5]. But then again, when applied on large collections, the automation is especially important in order to provide the scalability needed to annotate existing documents and reduce the burden of annotating new documents [6]. This automation is typically implemented with Information Extraction techniques, among which Named Entity Recognition is used to identify concepts to annotate.
- Question answering systems provide concrete answers to queries, and NER techniques are frequently used for this kind of systems as a means to facilitate the selection of answers. Indeed, in TREC-8 (Text REtrieval Conference) about 80% of the queries used to evaluate this kind of systems were of the type *who*, *where* and *when* [7], which use to be answered with named entities of type person, organization, localization and date.
- The Semantic Web and all the applications it supports depend on technology to make information interoperable, and ontologies play a key role in this ambitious scenario [8][9]. One of its cornerstones is therefore the proliferation of ontologies, which requires engineering for their quick and simple construction [10]. One of the tasks in this line is the automatic population of ontologies, which aims at incorporating instances into existing ontologies without the intervention of humans. Named Entity Recognition emerges in this case to identify instances of the required concepts. An example of this kind of systems can be found in the tool *KnowItAll* [11], which applies bootstrapping techniques upon the Web to achieve this goal.
- The Social Web has expanded in recent years, with people freely expressing their opinions on a variety of topics. When making a decision, more and more people look online for opinions related to whichever they are interested in, and they often base their final decision on the information found [12]. One of the pre-processing techniques applied for Opinion

Mining is the recognition of named entities of interest, from which related opinions can be identified and assessed as positive or negative (e.g. digital cameras). The system *OPINE* has been developed for the extraction of attributes of products and the analysis of the related opinions [13], and it is based on the tool *KnowItAll*, mentioned above.

All things considered, Named Entity Recognition is used for the identification of elements with certain semantics of interest, which can later be related one another to make up more complex scenarios.

3 Evolution of Named Entity Recognition

The term *Named Entity* was coined in 1996, at the 6th MUC conference, to refer to "unique identifiers of entities" [14]. However, already in 1991 Lisa F. Rau proposed a method for extracting company names from text, which is nowadays commonly accepted as a NER task. She realized that these entities were problematic for Natural Language Processing because they were mostly unknown words. Nonetheless, she described them as relevant pieces of information for topic analysis, for extraction of information from text, and for indexing a text for full-text retrieval [15].

In 2002 Petasis and colleagues limited the NE definition to "a proper noun, serving as a name for something or someone" [16]. They justified this restriction merely because of the significant percentage of proper nouns present in a corpus. On the other hand, Alfonseca and Manandhar defined NER as "the task of classifying unknown objects in known hierarchies that are of interest for us for being useful to solve a particular problem" [17]. This approach has been followed by the BBN hierarchy [18] and the GENIA ontology [19] for Information Retrieval (IR) and Question Answering (QA) tasks.

Another definition of NE was proposed in 2007 by Nadeau and colleagues. They claimed that "*named* aims to restrict the task to only those entities for which one or many rigid designators, as defined by Kripke, stand for the referent" [20][21]. Kripke's characterization of rigidity was proposed in 1980: "a designator *d* of an object *x* is rigid if it designates *x* with respect to all possible worlds where *x* exists, and never designates an object other than x with respect to any possible world" [22].

Despite the disagreement as to what a NE is, there has been some agreement as to the types of NE to recognize in the main NER evaluation forums. Between 1996 and 2008 the evaluation of NER tools was carried out in the MUC, CoNLL (Computational Natural Language Learning) [23] and ACE (Automatic Content Extraction) [24] conferences. The definition of the task was quite similar: given concrete types of semantics, the goal was to locate elements in the text that fit those semantics. Person, organization and localization were types of NE present in these forums, while temporal and numerical expressions were specific of MUC and ACE, and the miscellaneous type was unique of CoNLL. ACE also included types for facilities, weapons, vehicles and geo-politic entities, all of which were further broken down into several subtypes. The evaluation of NER tools was possible with the use of annotated corpora, generally small and from the journalistic domain for the most part.

Nowadays there are other conferences related to NER, though they tackle a different problem: the INEX Entity Ranking track (XER) [25], the TREC Entity Track (ET) [26] and the TAC Knowledge Base Population task (KBP) [27]. XER began in 2007 with the goal of ranking entities (Wikipedia articles) by their relevance to a question. ET started in 2009 to promote development of search engines capable of returning the homepages of the entities being searched, currently restricted to people, organizations, localizations and products. For instance, for the question "airlines that currently use Boeing 747 aircrafts", the returned results should be the homepages of airlines

meeting this requirement [28]. KBP began in the TAC conferences to complete attributes about given NEs, which are so far limited to people (e.g. alias, cause of death and date of birth), organizations and geo-politic localizations. In this case though, the task is more related with the identification of relations between entities and their properties than with the identification of the entities themselves.

As to the NER tools, several projects have been carried out in different areas and domains where new categories of NE have emerged. For example, names of genes and genetic products are treated in biological research [29]. In addition, commercial NER tools offer more and more predefined types of NE, besides the possibility of adapting to new types as requested by users. For instance, the *Calais* family of products, by Thomson Reuters, currently recognizes 39 different types of NE, among which we can find TV shows and sports leagues. *Thing Finder*, a commercial tool by Inxight, recognizes 45 predefined types of NER, among which there are some highly unusual ones such as holidays and financial indexes.

4 What is a Named Entity?

Experts in Named Entity Recognition have given several and different definitions of what a NE is. An analysis of these definitions allows us to categorize them in terms of the following four criteria: grammatical category, rigid designation, unique identification and domain of application. In order to determine what factors really define whether something is a NE or not, we selected several NEs from various sources and we analyzed them according to the above criteria (see Table 1).

Named Entity	Proper Noun	Rigid Desig.	Unique identifier	Example of domain/purpose	Source and NE type
water	No	Yes	? (sparkling or still)	Chemistry	SH: natur. obj-mineral
$10million	No	?	? (American dollars or Mexican pesos)	Finances	MUC-7: Money
whale	No	?	? (orca or minke)	Biology	SH: sea animal
Bedouins	No	?	? (specific people)	Army/ News	CoNLL03: Miscellanea
Airbus A310	Yes	?	? (specific airplane)	Army/Tech. Watch	CoNLL03: Miscellanea
Batman	Yes	Yes	? (people disguised)	Movies/Costumes	ACE: Person
Cytokine	Yes	?	? (specific proteins)	Biomedicine	GEN: protein family / group
twelve o'clock noon	No	?	? (specific day)	News	MUC-7: Time

Table 1. Examples of NE from Sekine's Hierarchy (SH) and MUC, CoNLL03, ACE and GENIA (GEN) annotated corpora or guidelines. Question marks indicate context-dependency.

4.1 Grammatical Category: Proper Nouns

Named Entities have been defined as proper nouns, or common names acting as proper nouns, because this grammatical category designates beings and unique realities with identifying function. Indeed, previous works state the problem of NER as the recognition of proper nouns in general. However, considering the objectives of the areas in which NER is applied, this does not seem enough to define what NEs are, as pointed out in the criteria adopted to tag NEs in the ANCORA corpus: "the classic grammatical approach to proper noun analysis is surely insufficient to deal with the problems NER poses" [2]. This is because some of the common characteristics of proper nouns, such as the lack of inflexions or determinants, the lack of lexical meaning and the use of capital

letters, are insufficient to describe named entities. We observe that some terms in Table 1 are not proper nouns, like *water* and *whale*. Therefore, this characteristic is easily rejected as a criterion to define NEs.

4.2 Rigid Designator

Several works on NER keep citing Kripke's definition of rigid designator from Theory of Names as an essential part of the NE definition, though they recognize that this definition is loosened sometimes for practical reasons [20]. Regarding the identity of rigid designators, Kripke gave the example of *Richard Nixon* as opposed to *President of the United States of America*, which is not a rigid designator because the referent element changes every few years. Also *Hesperus* and *Phosphorus* would be rigid designators, as both of them reference the planet Venus in every possible world where Venus exists, and they do not designate anything else in any other world. These concepts come from the Philosophy of Language, and they tend to be quite controversial [30]. For example, there is general agreement accepting *Phosphorus* and *Richard Nixon* as rigid designators, although nowadays *Phosphorus* also refers to the chemical element (indeed, named after the planet's glowing condition), and *Richard Nixon* might refer to many people with the same name, not just the former President. Following Kripke, even terms not usually considered as named entities, such as *gold*, *hot*, *redness* or *loudness*, are rigid designators, so the characteristic of rigid designation is not clear either. Except for cases where they have been explicitly designated as such, we cannot ensure that the examples in Table 1 meet this criterion.

4.3 Unique Identifier

Related to rigid designation, another way to define NEs found in the literature is with the concept of "unique identification". MUC conferences stated that the expressions to be annotated are "unique identifiers" of entities. However, the same entity with the same label might designate different referents (e.g. *Phosphorus*), and so what the "unique identifier" is considered is actually the referent of that to which we refer. This statement implies the previous knowledge of the referent in order to determine that it is unique and unambiguous, or contextual elements allowing it to be defined as such. Following this argument, virtually everything could be referred to uniquely, depending on the context or the previous knowledge of the receiver, although a unique identifier for one receiver might not be so for another one, either because of lack of shared knowledge or the ambiguity of the context. As shown in Table 1, we cannot ensure the existence of unique identifiers in all cases because it depends on the objective. For example, the classification of *Airbus A310* as unique identifier of a type of aircraft seems clear, but the classification of the concept *water* does not seem to respond to it being a unique identifier, unless we want to recognize different types of water, such as *sparkling water* and *still water*. On the other hand, not even proper nouns, usually considered as NEs, are unique identifiers. Their referent simply happens to be easier to locate (e.g. *Springfield* is easier to locate in a context-free discourse than *the center of the court*, despite there is ambiguity in both cases; the same occurs with H_2O and *water*, where the former is usually considered a NE and the latter is not).

4.4 Purpose and Domain of Application

The purpose and domain of application have determined the named entities to recognize from the beginning. In the MUC conferences, sponsored by DARPA, the objective of semantically classifying certain types of information and not others had a clear military application. Typical examples in this series of conferences were to determine the agent, time, cause and localization of an event off

some text. Another clear example is the inclusion of categories like vehicles or weapons in the ACE program.

Furthermore, the Question Answering area has influenced notably the definition of named entity. The NE hierarchy proposed by Sekine, with more than 200 categories (see Figure 1), justifies its extension because 7 or 8 classes are not enough to cover the needs of the applications, referring to QA and Information Extraction in general, more specifically in the journalism domain [31]. In fact, it is stated that it is possible to consider that NEs usually respond to the 5 typical questions in the journalism domain: *what*, *where*, *when*, *who* and *why*, which are capable of summarizing events and news. This characteristic has been exploited to capture this type of entities, thanks to an observed correlation between being an entity and appearing in a particular place in the text [32].

Therefore, the definition of NE according to the purpose and domain of application seems the only one consistent with the literature, evaluation forums and tools.

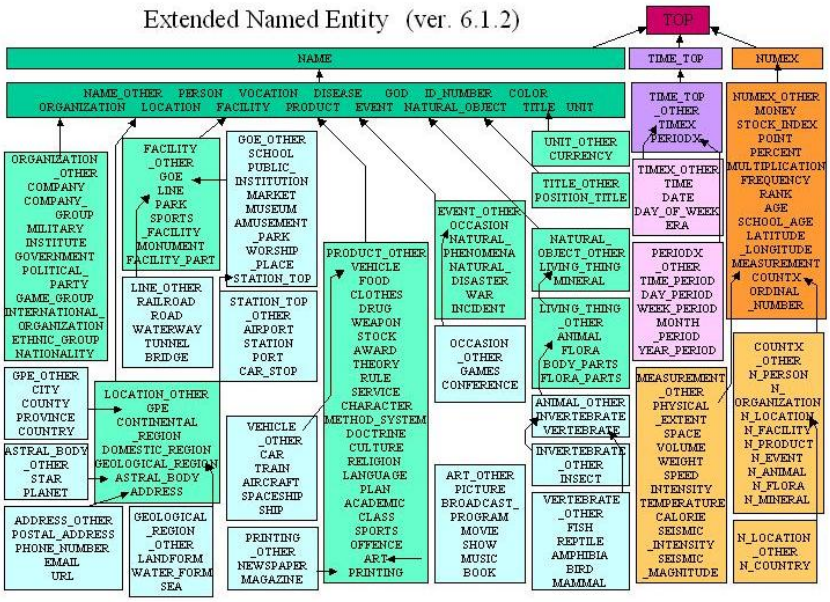

Figure 1. Sekine's Named Entity Hierarchy.

5 Consequences for NER Evaluation

Annotation criteria are different across the MUC, CoNLL and ACE evaluation forums. The most conflictive aspects in this matter are the types of Named Entity to recognize, the criteria for their identification and annotation, and their boundaries in the text. Such disagreements are shown in Table 2: we collected several entities from the corpora and annotation guidelines in MUC, CoNLL and ACE, and then annotated them following the other guidelines.

Types of Named Entity to identify. The types of NE considered in the international evaluation forums until 2008 agreed fairly well, though they were very limited (see Table 3). The wealth of types of NE considered nowadays, as well as the need to recognize new types suggested by the users, clearly expose how limited the traditional forums have been to evaluate meaningful types of NEs representative of the end user needs.

Identification and Annotation Criteria. The annotation guidelines followed in the traditional evaluation forums show some degree of confusion on how to annotate Named Entities. For instance, the guidelines used in ACE establish that in many situations it is confusing to differentiate

between names and other mentions (references to entities): *Health Ministry* is considered a reference (nominal mention), but *Ministry of Health* is considered a Named Entity (name mention) due to the number of words following the so called "trumping rule". Saints are not tagged at all in MUC "because removal of a saint's title leaves a non-unique name", although the removal of role names such as *President* are correct when tagging names of presidents. It is not clear either how to proceed with cases of metonymy (the use of a term in place of another related concept, as in *the throne* to refer to the monarchy). MUC's guide establishes that when it is common, the semantics of the term used shall be kept. According to this, in the sentence *Baltimore defeated the Yankees*, *Baltimore* is tagged as localization, while *Yankees* is considered a sports team (category contemplated as part of organization by CoNLL and ACE). In other cases the semantic meaning of the text is kept: airports become organizations instead of localizations if they refer to the airport's organization or business. Instead, ACE's guide suggests keeping the author's intentions in relation with the text.

Phrase	MUC-7	CoNLL03	ACE	Disag.
Baltimore defeated the Yankees	<Baltimore>LOC <Yankees>ORG (ref. A.1.6)	<Baltimore>ORG <Yankees>ORG	<Baltimore>NAM.ORG.SPO <Yankees>NAM.ORG.SPO (ref. 6.2)	C
Zywiec Full Light	<Zywiec>ORG ("Full Light" no markup, ref. A.1.7)	<Zywiec>ORG <Full Light>MISC	<Zywiec>NAM.ORG (ref. 9.3.2)	I, C
Empire State Building	no markup (ref. 4.2.3)	<Empire State>LOC	<Empire State Building> NAM.FAC.Building (ref. 9.3.2)	I, C, B
Alpine Skiing-Women's World Cup Downhill	no markup (ref. A.2.4)	<World Cup>MISC (ref. guide)	<Women>NOM <World>NOM (ref. 9.3.3)	I, C, B
the new upper house of Czech parliament	<parliament>ORG (ref. A.4.3, A.1.5)	<Czech>LOC	<Czech parliament>NOM (ref. 9.3.2)	I, C, B
Stalinist nations	no markup (ref. A.1.6)	<Stalinist>MISC	no markup (ref. 5.2.1)	I
Wall Street Journal	no markup (ref. A.1.7)	<Wall Street Journal>ORG	<Wall Street Journal> NAM.ORG.MED (ref. 9.5.3)	I

Table 2. Examples obtained from MUC, CoNLL03 and ACE. Comparative tagging with examples originally found in the corpora or annotation guidelines, and corresponding annotations following the other guidelines. The table shows the disagreements in identification of NE (I), boundaries (B) and assigned category (C).

Valid Boundaries of a Named Entity. The same Named Entities are sometimes annotated differently depending on the evaluation forum. For instance, in MUC and CoNLL it is common to annotate just the main word of a NE, ignoring terms it might have associated to itself as part of the same noun phrase (e.g. *Vice President*<*John Hime*>PER). Even when the word behaves as a qualifier, it is tagged as NE, ignoring the noun phrase's core (e.g. <*Bridgestone*>ORG *profits*, <*Clinton*>PER *government*, <*Kennedy*>PER *family*). It is even possible to find parts of compound words that have been annotated partially (e.g. <*Ford*>ORG *Taurus*), arguing that it is clear from context or annotator's knowledge that part of the whole word is an entity. In ACE the noun phrase and its whole semantic category is kept instead, with all mentions indicated with nested tags.

Conflicts as to the annotation criteria and boundaries of NEs make comparisons across forums unjustified. In addition, the results of NE tools are evaluated differently too. In MUC the identification is considered correct only if the boundaries of the NE are exactly as annotated, while the classification is considered correct regardless of the accuracy of the boundaries. In CoNLL a result is considered correct only if the boundaries and classification are exactly as annotated. In ACE, partial identifications are taken into account only if a certain proportion of characters from the main part of the NE are returned. Therefore, it is expected to observe very different performance

scores when following one or other annotation criteria, which is especially striking considering that none of them can be considered absolutely correct: there are not good or bad results in general, but good or bad results for a particular task definition and purpose.

MUC-7	CoNLL-03	ACE-08
Person	Person	Person (animal names excluded)
Organization	Organization	Organization
Localization	Localization	Localization / Geo-political name/ Facility
–	Miscellaneous	Person (as not structured group) / Vehicle / Weapon
Time / Date	–	Temporal expression (independent task)
Currency / Percentage	–	Quantity (independent task)

Table 3. Approximate relations among MUC7, CoNLL03 and ACE08 types of NE recognized.

6 Consequences for NER Tools

The ambiguity in the definition of Named Entity affects the NER tools too. We studied five research NER tools (*Annie, Afner, TextPro, YooName* and *Supersense Tagger* with the three learning models it incorporates) and two commercial tools (*ClearForest* and *LingPipe*), noting that the types of NE recognized are usually prefixed and very different across tools [33]. Apparently, they seem to implicitly agree on recognizing the categories of people, organization and localization as types of NE, but there are many discrepancies with the rest. Dates and numerical quantities are recognized as NEs by *Annie* (dates, currency and percentages), *Afner* (dates) and *YooName* (currency, dates and other units of measurement). Other categories such as food products and natural elements, or even names of events such as wars or hurricanes, are far less frequent and only *YooName* claims to recognize them (at least without the need of previous adaptation). Besides discrepancies across tools as to the boundaries of the NEs, we can go from recognizing at least 13 different types of semantics (not including subtypes) to just 3 types on the same corpus. As a result, it is unclear what to expect from a generic NER tool.

In some cases, the tools can be adapted to new types of NE through a learning process with an annotated corpus. This is the case of most supervised-learning tools, such as Supersense Tagger, which have several sets of tags and annotated corpora to train, obtaining totally different NEs as a result. The problem in this case is actually the lack of such corpora for arbitrary new types of NE, which would again be biased by the evaluation forums. Most importantly, if tools are not corpus-independent and they need large amounts of NEs annotated beforehand to be able to work, they have no practical use whatsoever for the average user.

In addition, it is very common to find types of NE easily identifiable by typographic characteristics, such as numbers and dates which, oddly, are not usually recognized when written with letters instead of digits. We may therefore find types of NE just because they are relatively easy to identify, not because they are reportedly useful. There are also clearly ambiguous categories, such as miscellaneous. *YooName* takes into account categories for currency, dates and measurements, but other tools either do not recognize them at all or just classify them in a category named miscellaneous, unknown or others. Without further classification, this category has no practical application because it implies that a NE is useful despite not knowing its concrete semantic category.

7 Is NER Really Solved?

The tools evaluated in MUC obtained very high precision and recall scores, beyond 90%. In particular, the best system obtained 97% precision and 96% recall [1]. A total of 16 systems participated in CoNLL03, where average ratios ranged between 60% and 90%. In ACE 2005 the scores where between 4% and 72%, with five of the systems scoring around 70%. In ACE 2008 though, the best scores were only marginally above 50%. However, the performance measures used in ACE are so different from the traditional precision-recall measures that scores from ACE are not at all comparable to scores from MUC and CoNLL. In any case, the fact that the best scores in 2008 were just about 50% indicates there is significant room for improvement in NER tools. Despite of this, in 2005 NER is generally regarded as a solved problem with performance scores above 95%, and after ACE 2008 these tasks are dropped from the international IE and IR evaluation forums.

Experimental validity establishes how well an experiment meets the well-grounded requirements of the scientific method, that is, whether the results obtained do fairly and actually assess what the experimenter attempted to measure [34][35]. The statement that NER is a solved problem is in fact based on results from several evaluation experiments, which are also subject to validity analysis [36]. For such a statement to be valid, those experiments must have met certain requirements. Next we analyze them in terms of content, external, convergent and conclusion validity.

7.1 Content Validity

Content validity evaluates the extent to which the experimental units reflect and represent the elements of the domain under study. In evaluation of NER tools, it requires the task to reflect the needs of the real users it is intended for. That is, the NER task has to define an adequate and meaningful set of semantics of interest, according to the user requirements, and a collection of documents representative as much as possible of the ones expected in real settings.

The semantics of interest in NER depend on the particular application. Goals can be very diverse, such as Ontology Population with tools such as *KnowItAll*, capable of identifying examples similar to others previously annotated thanks to the use of bootstrapping techniques over the Web. In the young area of Opinion Mining, NER tools can be used for the identification of products or other concepts of interest. These applications suggest the recognition of virtually any concept. General purpose NER tools also reflect this situation, moving from recognizing a few types of entities from the journalistic (people, organizations, organizations, dates) or military domains (vehicles, weapons), to new categories such as food products and names of events like hurricanes as in the tool *YooName*. Commercial tools offer more and more predefined types of NE with every new version, besides the possibility of adapting to new entities if required by the users.

This recent increase in the types of NE supposedly identifiable by the tools contrasts with the limitations of the NER evaluation forums up to 2008. Even if a tool performed very well with the limited number of types traditionally evaluated in these forums, it is completely unjustified to generalize the results and assume that they perform well with these new types of NE too, especially so considering that the corpora used back then were very limited too. Therefore, we find that with time there are clear needs for tools to identify new types of NE, and yet there is no generally accepted forum to evaluate how well they perform. This begs the question: do the tools that solved the NER problem back in 2005-2008 also solve the problem now in 2012?

7.2 External Validity

External validity evaluates the extent to which the results of an experiment can be generalized to other populations and experimental settings [36]. This is probably the weakest point in IR evaluation in general [37], mainly because the selection of documents and queries does not follow a random process. In fact, it has been shown that positive results in a test collection or an evaluation forum do not imply good performance with a different collection [37]. The approach traditionally followed consists in the use of very large corpora and the creation of different query sets from year to year, allowing researchers to compare their systems over time with different collections rather than just one.

Previous work in NER has observed variation in performance of NER tools when changing the density of NEs to identify, the way in which NEs are cited (e.g. a person is not cited the same way in the Wall Street Journal and the New York Times), the length of the documents, etc. [38]. In order to limit these negative effects it is important to use heterogeneous documents, which often means using large corpora. However, the NER corpora traditionally used are very small and biased towards the journalistic domain. The largest and most diverse was the one used in ACE 2005, with just about 50K words and different genres (Broadcast Conversation transcripts, Broadcast News transcripts, Conversational Telephone Speech transcripts, Meeting transcripts, Newswire, Usenet Newsgroup/Discussion Groups and Weblogs). An analysis of NER tools suggests a loss in performance when changing the corpus [33], and it has been shown that changing the sources, even maintaining the domain and changing just the genre, leads to significant performance losses between 20% and 40% [39]. This would drop the performance scores from 95% to about 65%, leveling NER with other Information Extraction tasks considered much difficult and not at all solved, such as Scenario Template Production.

7.3 Convergent Validity

Convergent validity evaluates the extent to which the results of an experiment agree with other results, theoretical or experimental, they should be related with [36]. The validity of IR evaluation experiments is sometimes questioned because the results obtained use to vary depending on the actual people that make relevance judgments. By considering with how well a human performs when compared to another human, agreement scores between annotators provide us with a sense of how well a system is expected to perform. For instance, if the agreement between human annotators were 100%, an automatic system can be expected to return perfect results; but if the agreement was just 50%, then we can hardly expect a system to perform better than 50%. If it did, then it would provide exceptionally good results for the particular annotator, but for an arbitrary person it would be expected to perform much worse.

The agreement scores between annotators depend on three main factors: previous knowledge on the domain, the annotation guidelines, and the semantics to annotate and collection used. Reported agreement scores in the Biomedicine domain are very variable, with 75% to 90% in the annotation of genes and proteins [40]. In the case of ACE, agreement scores were reported to be above 80%. In particular, the scores in the 2002 EDT task (Entity Detection and Tracking) were 86%, up to 88% in 2003. These agreement scores are similar to the ones reported for MUC [41]. The systems evaluated in ACE up to 2008 were very far from these scores, but in MUC they generally scored above 90% in F-measure, getting even to 96%. To the best of our knowledge, there is no task in Information Retrieval where systems generally outperform the human agreement scores, which could be explained by systems overfitting to the training corpora. As we saw above, this leads to serious limitations in the external validity of the NER evaluations, where significant

loss of effectiveness is expected and in fact observed: systems do not perform over 95% in general but for a particular user and document type.

7.4 Conclusion Validity

Conclusion validity evaluates the extent to which the conclusions drawn from the results of an experiment are justified [36]. In fact, any unhandled threat to validity in the evaluations may lead us to invalid overall conclusions. In the case of NER, the problems mentioned above show clear limitations in terms of experimental validity of the evaluation forums, effectively invalidating the statement that NER is a solved problem.

In addition, each of the three main forums used different criteria and measures to assess the effectiveness of tools, complicating the comparison of results even when dealing with similar semantics and corpora. ACE is the only forum with a trajectory long enough so as to measure the evolution of the area and draw general conclusions, but its overly complicated measures are so hard to interpret that it is very difficult to assess the state of the art and how the improvements, if there were any, translate into real operational settings. Therefore, we believe there is not enough evidence to support the statement that NER is solved: it rather suggests the opposite.

8 Current Evaluation Forums

The traditional NER evaluation forums stepped aside and were followed by the new XER and ET tasks, which significantly extended both the number of semantics to capture and the document collection. But these new forums do not replace MUC, CoNLL and ACE because the concept of NE is reduced to a document. In fact, in ET they adopted the definition of NE as "things represented by their homepages on the web" [42]. The retrieval units are thus perfectly defined and it is possible to apply evaluation methods similar to the ones applied in general IR tasks and widely studied in TREC. The problem lies in that the techniques employed to locate a document do not need to be similar to the ones employed in identifying and delimiting specific parts of text, necessary for all applications of NER. As with Question Answering systems, answer strings may be supported by documents, but the retrieval units are still not defined and part of the problem is actually deciding where the boundaries of the answer text are.

Another problem is that the effectiveness of NER tools is measured not only by precision, but also by recall. A NER tool used for semantic annotation or ontology population, for instance, is expected to identify all elements that meet the concrete semantics. IR search engines, on the other hand, are not expected to retrieve all relevant information, but rather to retrieve it early in the results list and, if possible, to retrieve a perfect result at the top of the list. As a consequence, the measures used in the new forums are often focused on precision and the ranking of results, but not on recall [43]. In the traditional NER task there is no such thing as a ranking of entities or a perfect result, so using these measures leads us to a problem of construct validity because the results of the evaluation experiments do not measure what we intend to [36].

Finally, we have to bear in mind that not all identifiable semantics have representative Web homepages (e.g. monetary quantities or dates), and neither do all possible entities of a particular semantics, no matter how common it seems (e.g. not all businesses have a Web page).

In summary, and considering the applications of NER, we need techniques and tools capable of recognizing semantics of interest in free text and that do not necessarily require the existence of full documents like Web pages or Wikipedia articles that support them. Therefore, the new NER-related evaluation forums cannot be considered just a modern version of MUC, CoNLL and ACE,

because they deal with very different tasks. Any link between the results of these forums and the traditional NER task should thus be made with care.

9 Challenges and Opportunities in NER

NER is not a solved task, but it can be solved. At least, to the extent any other domain-dependent task can be considered as solved. The problem is that current evaluation practices and resources in NER do not allow us to decide. NER has been considered a solved problem when the techniques achieved a minimum performance with a handful of NE types, document genre and usually in the journalistic domain. We do not know how well current techniques perform with other types of NE and different kinds of documents. There are no commonly accepted resources to evaluate the new types of NE that tools recognize nowadays, and the new evaluation forums, though they overcome some of the previous limitations, are not enough to measure the evolution of NER because they evaluate systems with different goals, not valid for most NER applications.

Therefore, the NER community is presented with the opportunity to further advance in the recognition of any named entity type within any kind of collection. But any attempt to evaluate such generic NER tools for arbitrary needs will probably fail because of the enormous effort it would require. Instead, evaluation forums should focus on specific applications that allow researchers to focus on specific user needs and progressively improve NER techniques. In particular, it is necessary to extend the typology of NEs and vary it with some frequency, lining up with current user needs. The evaluation corpora need to be extended too, paying special attention to the application of NER in such heterogeneous scenarios like the Web. These evaluation forums should be maintained along time, with stable measures, agreed upon and shared by the whole community. This would allow us to measure the actual domain-specific state of the art and, at some point, have at our disposal a sufficiently large and generic evaluation framework to assess whether the generic NER task is a solved problem or not. In the mean time, we should only address this question for specific applications.

But getting to adequate NER evaluations without raising costs is a challenging problem. The effectiveness measures widely used in other areas are not suitable for NER, and the evaluation methodologies need to be reconsidered. In particular, it is necessary to measure recall, which can be extremely costly with a large corpus. An alternative is to attempt to identify the existing NEs with the aid of various NER tools and incrementally evaluate and add the new NEs identified by the participating systems. Although this still does not allow us to accurately measure exact recall, making the collection and results publicly available would allow researchers to compare with any other system and reduce this problem. This method has already been used in the TREC-QA list task to recognize answers in the form of lists, which could be considered as the basis to evaluate NER tools given the similarities.

Finally, it is necessary to reconsider the effort required to adapt a tool to a new type of entity or collection, as it usually implies the annotation of a new document collection. The recurrent use of supervised machine learning techniques during the last decade contributed in making these tools portable, but at the expense of significant annotation efforts on behalf of the end user. A tool requiring an annotated corpus in the order of hundreds of megabytes is not comparable with, and should not be compared to, another one that requires the annotation of just a handful examples suggested by the tool itself. The real applicability of the former largely depends on the resources available to the end user, and therefore its evaluation should contemplate how much effort is required on her behalf to adapt the tool to her needs. Evaluations in NER must include not only effectiveness measures, but also cost measures that assess how much effort these tools would

require from the end user. These measures would promote research on low-cost methodologies capable of adapting to new entities in new domains, reducing the annotation cost for the user. Bootstrapping and active learning techniques are examples in this line. Furthermore, the development of this kind of tools would contribute to the creation of the required resources to perform the dynamic and continued evaluations that are necessary to properly measure the evolution of the field.

Therefore, and in spite of the linguistic conflict regarding the Named Entity term and the derived problems for the tools and evaluation forums, advancing in the NER task is feasible. As in other similar areas, its advancement should be progressive. Here we do not imply the need for generic NER tools, but the need to advance in techniques and resources that prove useful for different applications and that, eventually, could lead to a significant degree of portability. This portability is even more important if we consider that any semantic of interest for a user, within a domain and for a particular purpose, can indeed be considered a named entity.

10 Conclusions

Named Entity Recognition plays a very important role in other Information Extraction tasks such as Identification of Relationships and Scenario Template Production, as well as other areas such as Semantic Annotation, Ontology Population or Opinion Mining, just to name a few. However, the definitions given for Named Entity have been very diverse, ambiguous and incongruent so far.

The evaluation of NER tools has been carried out in several forums, and it is generally considered a solved problem with very high performance ratios. But these evaluations have used a very limited set of NE types that has seldom changed over the years, and extremely small corpora compared to other areas of Information Retrieval. Both factors seem to lead to overfitting of tools to these corpora, limiting the evolution of the area and leading to wrong conclusions when generalizing the results. It is necessary to take NER back to the research community and develop adequate evaluation forums, with a clear definition of the task and user models, and the use of appropriate measures and standard methodologies. Only by doing so may we really contemplate the possibility of NER being a solved problem.

References

[1] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *16th Conference on Computational linguistics*, 1996, pp. 466-471.

[2] O. Borrega, M. Taulé, and M. A. Martí, "What do we mean when we speak about Named Entities," in *Conference on Corpus Linguistics*, 2007.

[3] H. Cunningham, "Information extraction, automatic," in *Encyclopedia of Language and Linguistics*, 2nd ed., Elsevier, 2005, pp. 665-677.

[4] J. Gantz and D. Reinsel, "The Digital Universe Decade, Are You Ready?," 2010.

[5] V. Uren et al., "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Journal of Web Semantics*, vol. 4, no. 1, pp. 14-28, 2006.

[6] L. Reeve and H. Han, "Survey of semantic annotation platforms," *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05*. ACM Press, New York, New York, USA, p. 1634, 2005.

[7] R. Srihari and W. Li, "Information Extraction Supported Question Answering," in *8th Text REtrieval Conference (TREC-8)*, 2000, no. 500, pp. 185-196.

[8] L. M. Á. Sabucedo, L. E. A. Rifón, R. M. Pérez, and J. M. S. Gago, "Providing standard-oriented data models and interfaces to eGovernment services: A semantic-driven approach," *Computer Standards & Interfaces*, vol. 31, no. 5, pp. 1014-1027, 2009.

[9] H. N. Talantikitea, D. Aissanib, and N. Boudjlidac, "Semantic annotations for web services discovery and composition," *Computer Standards & Interfaces*, vol. 31, no. 6, pp. 1108-1117, 2009.
[10] A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72-79, 2001.
[11] O. Etzioni et al., "Unsupervised named-entity extraction from the Web: An experimental study," *Artificial intelligence.*, vol. 165, no. 1, p. 91, 2005.
[12] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.
[13] A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of HLT/EMNLP*, 2005, vol. 5, pp. 339–346.
[14] N. Chinchor and P. Robinson, "MUC-7 named entity task definition," in *7th Conference on Message Understanding*, 1997.
[15] L. F. Rau, "Extracting Company Names from Text," in *7ª Conference on Artificial Intelligence Applications*, 1991, pp. 29-32.
[16] G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods," in *23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 128-135.
[17] E. Alfonseca and S. Manandhar, "An unsupervised method for general named entity recognition and automated concept discovery," in *1st International Conference on General WordNet*, 2002.
[18] A. Brunstein, "Annotation guidelines for answer types," 2002.
[19] Jd. Kim, T. Ohta, Y. Teteisi, and J. Tsujii, "GENIA corpus: a semantically annotated corpus for bio-textmining," *Bioinformatics (Oxford, England)*, vol. 19, no. 1, p. 180, 2003.
[20] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 7, 2007.
[21] D. Nadeau, "Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision," Department of Information Technology and Engineering, University of Ottawa, 2007.
[22] S. A. Kripke, *Naming and Necessity*. Harvard University Press, 1980.
[23] E. F. T. K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *CoNLL-2003*, 2003, pp. 142-147.
[24] NIST, "Automatic Content Extraction Evaluation (ACE08). Official Results.," 2008.
[25] G. Demartini, T. Iofciu, and A. D. Vries, "Overview of the INEX 2009 Entity Ranking Track," in *INEX 2009 Workshop Pre-proceedings*, S. Geva, J. Kamps, and A. Trotman, Eds. Amsterdam: IR Publications, 2009, pp. 233-237.
[26] K. Balog, P. Serdyukov and A. D. Vries, "Overview of the TREC 2010 Entity Track," in *Text REtrieval Conference*, 2010.
[27] H. Ji and R. Grishman, "Knowledge Base Population: Successful Approaches and Challenges," in *Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 1148–1158.
[28] K. Balog, P. Serdyukov, and A. P. de Vries, "Overview of the TREC 2010 Entity Track," in *Text REtrieval Conference*, 2010.
[29] M. Marrero, S. Sanchez-Cuadrado, J. Urbano, J. Morato, and J. A. Moreiro, "Information Retrieval Systems Adapted to the Biomedical Domain," ACM Computing Research Repository, 2012.
[30] J. LaPorte, "Rigid Designators," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. 2011.
[31] S. Sekine and C. Nobata, "Definition, dictionaries and tagger for extended named entity hierarchy," in *Language Resources and Evaluation Conference (LREC)*, 2004, pp. 1977-1980.
[32] Y. Shinyama and S. Sekine, "Named entity discovery using comparable news articles," in *International Conference on Computational Linguistics (COLING)*, 2004, pp. 848-853.
[33] M. Marrero, S. Sánchez-Cuadrado, J. Morato Lara, and Y. Andreadakis, "Evaluation of Named Entity Extraction Systems," in *10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'09)*, 2009.
[34] J. Katzer, K. H. Cook, and W. W. Crouch, *Evaluating Information: A Guide for Users of Social Science Research*, 4th ed. McGraw-Hill, 1998.
[35] M. L. Mitchell and J. M. Jolley, *Research Design Explained*, 7th ed. Wadsworth Publishing, 2009.

[36] J. Urbano, “Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain,” in *International Society for Music Information Retrieval Conference*, 2011, pp. 609-614.
[37] E. M. Voorhees, “The Philosophy of Information Retrieval Evaluation,” in *Workshop of the Cross-Language Evaluation Forum*, 2002, pp. 355-370.
[38] M. Vilain, J. Su, and S. Lubar, “Entity extraction is a boring solved problem: or is it?,” in *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, 2007, pp. 181-184.
[39] T. Poibeau and L. Kosseim, “Proper Name Extraction from Non-Journalistic Texts,” *Language and Computers*, vol. 37, pp. 144-157, 2001.
[40] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willet, “Protein structures and information extraction from biological texts: the Pasta system,” *Bioinformatics*, vol. 19, no. 1, pp. 135-143, 2003.
[41] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, “The Automatic Content Extraction (ACE) Program -Tasks , Data , and Evaluation,” in Conference on Language Resources and Evaluation, 2004.
[42] K. Balog, A. D. Vries, P. Serdyukov, P. Thomas, and T. Westerveld, “Overview of the TREC 2009 Entity Track,” in *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.
[43] J. Zobel, A. Moffat, and L. A. F. Park, “Against Recall: Is it Persistence, Cardinality, Density, Coverage, or Totality?,” *ACM SIGIR Forum*, 2009.