Natural language processing and it’s role in text analytics

Monica marrero

November 10, 2023

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 far from solved and show the consequences for the development and evaluation of tools. We discuss topics for further research to bring the task back to the research scenario.

1 Introduction

Named Entity Recognition (NER) is a task in Information Extraction consisting of 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)[7], where the importance of the semantic identification of people, organizations, and localizations, as well as numerical expressions such as time and quantities was clear. Most of the NER tools nowadays keep considering these types of NE which originated in MUC, though with important variations. however, it is unclear what the meaning of NE is This question has not been analyzed in 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 percent[4]. In this paper, we argue that NER is 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 us to reconsider whether NER is a solved problem or not, design evaluation procedures suitable for the current needs and 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 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 Unlocking the potential of technology, one keystroke at a time

According to a market survey performed by IDC[6], 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 the 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: ”In the bustling heart of the city, where the urban skyline meets the infinite sky, dreams take flight. Among the constant hum of life, stories unfold in every corner, and aspirations weave through the very fabric of existence. In this vibrant tapestry, individuals carve their paths, fueled by ambition and resilience. Amidst challenges and triumphs, they find the strength to rise above, creating moments that echo in the corridors of time. Each day brings new opportunities, each encounter is a chance for connection. It's a world where passion meets purpose, where the extraordinary resides in the ordinary, and where every heartbeat narrates a unique tale, waiting to be heard." Simply use the section and subsection commands, as in this example document! With Overleaf, all the formatting and numbering is handled automatically according to the template you've chosen. If you're using the Visual Editor, you can also create new sections and subsections via the buttons in the editor toolbar. In various domains, Named Entity Recognition (NER) plays a crucial role: 1. Information Retrieval and Interoperability: NER aids in improving information retrieval and enhancing interoperability between different data sources. It helps annotate documents, making them more accessible and compatible[16].Automation is especially vital for large document collections, ensuring scalability and reducing the burden of manual annotation[13].

2. Question-Answering Systems: NER techniques are frequently employed in question-answering systems to provide concrete answers to user queries. These systems often rely on identifying named entities like persons, organizations, localizations, and dates to facilitate accurate answer selection[14].

3.Semantic Web and Ontology Population: The Semantic Web aims to make information interoperable, and ontologies are central to achieving this goal. Automated ontology construction is vital, and NER assists in identifying instances of concepts for the ontology population. Tools like KnowItAll use bootstrapping techniques to automatically populate ontologies from web data[15][8][5][9].

4. Opinion Mining on the Social Web: With the growth of the social web, people express their opinions online. Opinion mining involves analyzing these opinions, and NER plays a role in identifying named entities of interest. For example, in product reviews, recognizing named entities like "digital cameras" helps identify and assess related opinions, whether they are positive or negative[?]. In these diverse applications, NER contributes to information organization, retrieval, and analysis, ultimately enhancing the efficiency and effectiveness of various information processing tasks.[11].

In these diverse applications, NER contributes to information organization, retrieval, and analysis, ultimately enhancing the efficiency and effectiveness of various information processing tasks.

2.1 Evolution of Named Entity Recognition

The term "Named Entity" was coined in 1996 at the 6th MUC conference, referring to "unique identifiers of entities"[4]. However, in 1991, Lisa F. Rau proposed a method for extracting company names from text, now accepted as an NER task, recognizing their importance for topic analysis, information extraction, and text indexing[12].

In 2002, Petasis and colleagues defined NE as "a proper noun, serving as a name for something or someone"[10]. Alfonseca and Manandhar defined NER as "the task of classifying unknown objects in known hierarchies"[1], followed by the BBN hierarchy[3] and the GENIA ontology[**?**] for IR and QA tasks.

Another NE definition in 2007 by Nadeau and colleagues emphasized "rigid designators" based on Kripke's work, characterizing rigidity in naming[**?**].

Despite these disagreements, NER evaluation forums between 1996 and 2008 (MUC, CoNLL, ACE) focused on locating elements in the text that fit predefined semantics. Common NE types included persons, organizations, and localizations. Additional types such as temporal and numerical expressions were specific to certain conferences. Recent NER-related conferences include the INEX entity ranking track (XER), TREC entity track (ET), and knowledge base population task (KBP). XER ranks entities' relevance to questions, ET focuses on returning homepages of searched entities, and KBP completes attributes about NEs, emphasizing relations between entities and their properties. NER tools have evolved, covering diverse domains and emerging NE categories. Commercial tools like Thomson Reuters' Calais and Inxight's Thing Finder recognize numerous predefined NE types, offering adaptability to user-requested types. For instance, Calais recognizes 39 NE types, including TV shows and sports leagues, while ThingFinder identifies 45 types, including holidays and financial indexes.

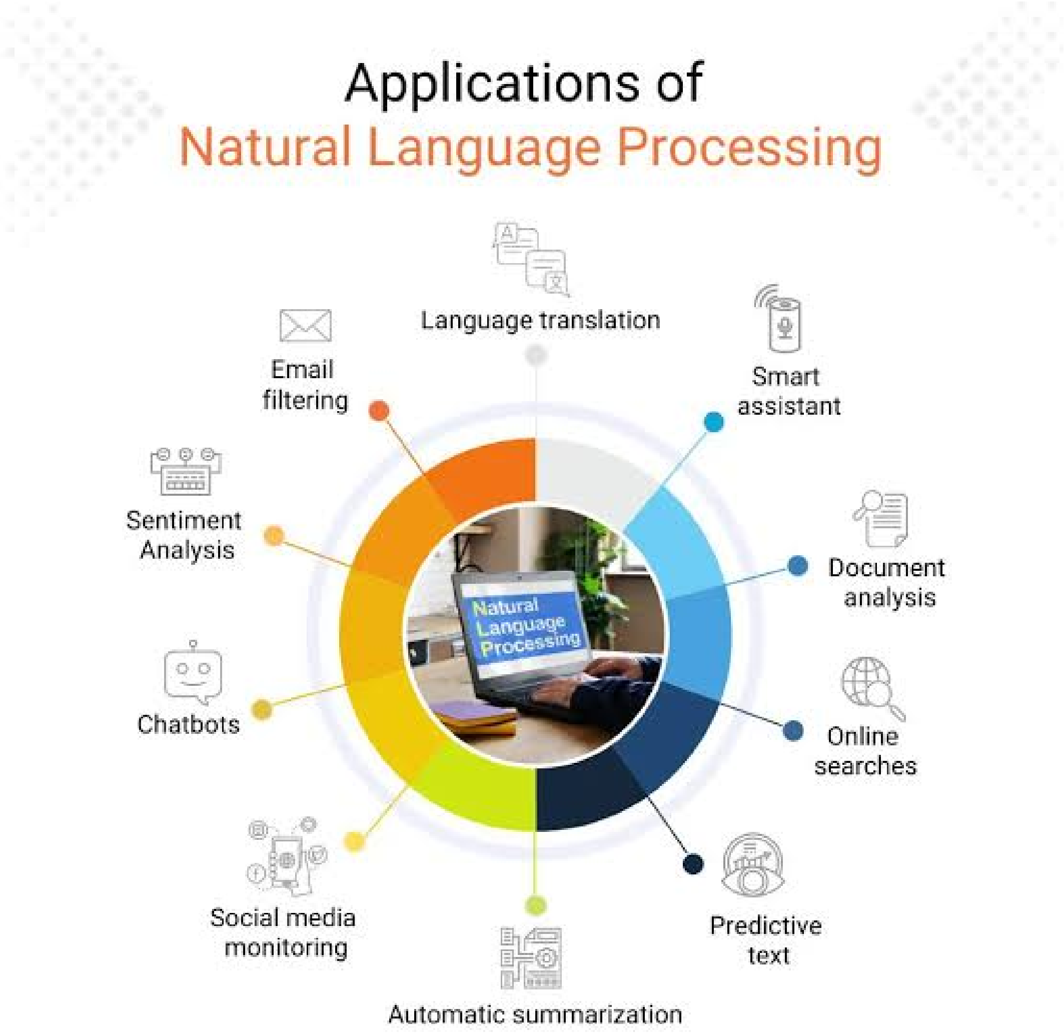

Figure 1: This image1 was uploaded via the file-tree menu.

Named Entity	Proper noun	Unique identifier	Example of domain/purpose
Water	No	? (Sparkling or still)	Chemistry
$10 million	No	? (American dollars or Mexican pesos)	Finances
Whale	No	? (Orca or minke)	Biology
Bedouins	No	? (Specific people)	Army/ News
Airbus A310	Yes	? (Specific airplane)	Army/Tech. Watch
Batman	Yes	? (People disguised)	Movies/Costumes
Cytokine	Yes	? (Specific proteins)	Biomedicine
Twelve o'clock noon	No	? (Specific day)	News

Table 1: Examples of NE from Sekine's hierarchy (SH) and MUC, CoNLL03, ACE and GENIA (GEN) annotated corpora or guidelines.

2.2 Consequences for NER evaluation

The challenges in Named Entity Recognition (NER) extend to defining valid boundaries for entities. Different valuation forums, such as MUc, CoNLL, and ACE, employ diverse criteria for annotation. MUC and CoNLL often annotate just the main word of an NE, disregarding associated terms within the same noun phrase. Even qualifiers or parts of compound words might be tagged as NEs. In contrast, ACE retains entire noun phrases with nested tags, maintaining the semantic category as a whole. These inconsistencies in annotation criteria hinder fair comparisons across forums. Moreover, discrepancies exist in evaluating NER tool results. In MUC, correctness is determined by exact boundaries and regardless of classification accuracy. CoNLL demands precise boundaries and classification for correctness. ACE, on the other hand, considers partial identifications if a certain proportion of characters from the main part of the NE is returned. These varying evaluation methods yield markedly different performance scores. This complexity underscores the importance of aligning evaluation criteria with specific task definitions and purposes, emphasizing the context-driven nature of NER assessments.

2.3 Consequences for NER tools

The lack of a clear, standardized definition for Named Entities (NE) has significant implications for NER tools. Analysis of various research and commercial tools like Annie, Afner, TextPro, YooName, ClearForest, and LingPipe reveals inconsistencies in recognized NE categories. While there's a consensus on entities like people, organizations, and locations, discrepancies arise in recognizing other types, including dates, numerical quantities, and specific events. Adaptation to new NE types is possible through supervised learning tools like Supersense Tagger, but this approach relies heavily on annotated corpora, often limited by biases from evaluation forums. Moreover, tools often struggle with corpus independence and practical usability, especially for the average user. Some NES, identifiable via typographic characteristics, are recognized, yet this recognition might not align with actual usefulness. Ambiguous categories like "miscellaneous" further complicate the practical application of NE tools, highlighting the challenges arising from the lack of standardized definitions.

3 Analysis

3.1 Current evaluation forums

The shift from traditional NER evaluation forums like MUC, CoNLL, and ACE to newer tasks like XER and ET has brought challenges. In these new forums, NEs are reduced to documents, deviating from the traditional approach where specific parts of text are identified and delimited. Furthermore, defining NEs solely as "things represented by their web homepages" in ET complicates the boundary determination problem. The evaluation measures, emphasizing precision and ranking, diverge significantly from traditional NER tasks where there's no perfect or ranked entity. Not all identifiable semantics have representative web pages, making these measures inadequate for diverse applications. Considering the wide-ranging applications of NER, tools must recognize semantics in free text without relying on the existence of full supporting documents. Therefore, the new evaluation forums, focusing on web-based entities, cannot be seamlessly compared with traditional NER tasks, demanding careful consideration of their applicability and relevance.

4 . Challenges and opportunities in NER

Named Entity Recognition (NER) remains an unsolved task due to challenges in evaluating its performance across diverse entity types and document genres. Historical evaluations, often limited to a few NE types in journalistic texts, do not reflect the broader applicability of NER techniques. New NE types recognized by tools lack comonly accepted evaluation resources. To advance NER, specific application-focused evaluation forums are needed, expanding NE typologies and incorporating diverse corpora, including web data. Efforts should also consider recall measurement, potentially through collaborative identification of NEs by multiple NER tools. Additionally, evaluations should address the effort required to adapt tools to new entity types, considering both effectiveness and cost measures. The field's evolution requires low-cost methodologies and resources for continuous evaluations. In conclusion, despite linguistic conflicts and ambiguities in NER definitions, progress is feasible through incremental advancements in techniques and resources, ultimately leading to greater tool portability. NER's importance extends to various information extraction tasks, semantic annotation, ontology population, and opinion mining, making its development crucial for a wide range of applications.

5 Conclusion

Named Entity Recognition plays a very important role in other Information Extraction tasks such as the 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 Entites 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. however 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 the 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 contemplate the possibility of NER being a solved problem

6 References

- An unsupervised method for general named entity recognition and automated concept discovery.In Proceedings of the 1st international conference on general WordNet, Mysore, India, pages 34–43, 2002. -Enrique Alfonseca and Suresh Manandhar

- hat do we mean when we speak about named entities. In Proceedings of Corpus Linguistics. Citeseer, 2007. -Oriol Borrega, Mariona Taule, and M Antonia Marti.

- Annotation guidelines for answer types. LDC2005T33, Linguistic Data Consortium, Philadel- phia, 2002 -Ada Brunstein

- Information extraction, automatic. Encyclopedia of language and linguistics,, 3(8):10, 2005 -Hamish Cunningham

- Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1):91–134, 2005 Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates.

- The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012):1–16, 2012 -john Gantz and David Reinsel

- Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996. -Ralph Grishman and Beth M Sundheim.

- Ontology learning for the semantic web. IEEE Intelligent systems, 16(2):72–79, 2001 -Alexander Maedche and Steffen Staab.

- Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2):1–135, 2008. -Bo Pang, Lillian Lee, et al

- Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 128–135, 2000 -Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis, and Con- stantine D Spyropoulos.

- Extracting product features and opinions from reviews. Natural language processing and text mining, pages 9–28, 2007 -Ana-Maria Popescu and Orena Etzioni

- Extracting company names from text. In Proceedings the Seventh IEEE Conference on Artificial Intelligence Application, pages 29–30. IEEE Computer Society, 1991. -Lisa F Rau

- Survey of semantic annotation platforms. In Proceedings of the 2005 ACM symposium on Applied computing, pages 1634–1638, 2005 -Lawrence Reeve and Hyoil Han

- Providing standard-oriented data models and interfaces to the government services: A semantic-driven approach. Computer Standards & Interfaces, 31(5):1014–1027, 2009

  Luis M óAlvarez Sabucedo, Luis E Anido Rif óon, Rub óen M óıguez P óerez, and Juan M Santos Gago.

- Computer Standards & Interfaces, 31(6):1108–1117, 2009. -Hassina Nacer Talantikite, Djamil Aissani, and Nacer Boudjlida. Semantic annotations for web services dis- covery and composition

- Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics, 4(1):14–28, 2006. -Victoria Uren, Philipp Cimiano, Jos óe Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna.

References

[1] Enrique Alfonseca and Suresh Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet, Mysore, India*, pages 34–43, 2002.

[2] Oriol Borrega, Mariona Taulé, and M Antø'nia Martı. What do we mean when we speak about named entities. In *Proceedings of Corpus Linguistics*. Citeseer, 2007.

[3] Ada Brunstein. Annotation guidelines for answer types. *LDC2005T33, Linguistic Data Consortium, Philadelphia*, 2002.

[4] Hamish Cunningham. Information extraction, automatic. *Encyclopedia of language and linguistics,*, 3(8):10, 2005.

[5] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134, 2005.

[6] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16, 2012.

[7] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[8] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79, 2001.

[9] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[10] Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–135, 2000.

[11] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. *Natural language processing and text mining*, pages 9–28, 2007.

[12] Lisa F Rau. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society, 1991.

[13] Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1634–1638, 2005.

[14] Luis M Álvarez Sabucedo, Luis E Anido Rifón, Rubén Míguez Pérez, and Juan M Santos Gago. Providing standard-oriented data models and interfaces to egovernment services: A semantic-driven approach. *Computer Standards & Interfaces*, 31(5):1014–1027, 2009.

[15] Hassina Nacer Talantikite, Djamil Aissani, and Nacer Boudjlida. Semantic annotations for web services discovery and composition. *Computer Standards & Interfaces*, 31(6):1108–1117, 2009.

[16] Victoria Uren, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4(1):14–28, 2006.