The history of named entity recognition

Named Entity Recognition & Geo-Tagging for Literary Analysis

Diana Santos

d.s.m.santos@ilos.uio.no

Distant Reading

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

COST action CA16204 is supported by the EU Framework Programme Horizon 2020

22 March 2021

Contents

- A simple definition – to be improved during the presentation
- MUC(K) (1987-1998)
- IREX (1998)
- ACE (2002-2008)
- CoNLL (2002; 2003)
- TimeML (2003)
- ENE (2004)
- HAREM (2006; 2008)
- TempEval (2007; 2010; 2013)
- SHINRA (2020)
- NER in literature
- The situation in COST (seen by the NER subgroup)
- A more informed task definition

An important evaluation paradigm

The evaluation contest, later on dubbed shared task, is a particular task/problem that several people want to address, and that implies:

1. agree on the problem
2. agree on how to evaluate the results
3. (often) prepare correct results (a golden collection)
4. (often) provide training materials
5. let a lot of participants try

MUC(K) - The message understanding conference

- The overall goal was to advance the performance of message understanding. See Hirschmann (1998). What we now call information extraction, or non-structured to structured conversion.
- MUC happened from 1987 to 1998 (MUC-7)
- MUCK used 10 messages from a tactical Navy domain
- MUCK-II incorporated a template

HOSTILE SURFACE FORCES IN VIC OF SEATTLE CONSIST OF ONE KRESTA I AND TWO KRIVAKS. 1115(L) USS CONOLLY FIRED FOUR HARPOONS AGAINST THE KRESTA I. NO BDA AT THIS TIME.

EVENT	ATTACK
INITIATING FORCE	FRIENDLY
AGENT CATEGORY	SURF
OBJECT CATEGORY	SURF
ID AGENT	USS CONOLLY DD-979
ID OBJECT	UR KRESTA I CLASS CV
INSTRUMENT/AGENT	HARPOON
LOC OF OBJECT	"IN VIC OF SEATTLE"
TIME	"1115(L)"
RESULT	NO DATA

MUC(K) - The message understanding conference (cont.d)

- MUC-3 started collective data gathering; MUC-5 added Japanese and two different domains: joint ventures and micro-electronics

General Electric Co. of the *U.S.* and *Taiwan's MITAC International Corp.* announced a joint venture to produce *advanced electronics* and information-processing equipment for defense use in *Taiwan*.

The partners will have equal stakes in the new company, dubbed *GETAC*. GE will take charge of technology transfer and foreign sales, while MITAC will be responsible for local management and sales, said Matthew Miau, MITAC Group chairman.

The venture will help improve the quality of Taiwan-made military products and promote the country's long-term goal of becoming a world-wide arms supplier, Mr. Miau said.

The partners expect to invest a total of $6.6 million in the project in the first year and to increase their investment to $14 million within the next five years.

TEMPLATE
DATE 100889
SOURCE Wall Street Journal
TIE-UP RELATION
ENTITY 1: NAME: General Electric Co
LOCATION: US
ENTITY 2: NAME: MITAC International Corp.
LOCATION: Taiwan
NEW-ENTITY: NAME: GETAC
INDUSTRY: Electronics

MUC(K) - The message understanding conference (cont.d)

- 3 (4) domain-independent tasks were added in MUC-7 (8)
  1. NE: persons, organizations, locations, times, dates, and money
  2. Co-reference
  3. Template element
  4. Template relation

General Electric Co. of the *U.S* and *Taiwan's MITAC International Corp.* announced a joint venture to produce advanced electronics and information-processing equipment for defense use in *Taiwan*.

The partners will have equal stakes in *the new company*, dubbed *GETAC*. *GE* will take charge of technology transfer and foreign sales, while *MITAC* will be responsible for local management and sales, said *Matthew Miau, MITAC Group chairman*.

ENTITY: Type ORGANIZATION
NAME: General Electric Co., GE
Category: ORG_CO
ENTITY: Type ORGANIZATION
NAME: MITAC, MITAC Group, MITAC International Corp.
Category: ORG_CO
ENTITY: Type ORGANIZATION
NAME: GETAC
Category: ORG_CO
Descr: the new company
PERSON: Type PERSON
Person: Matthew Miau
Category: PER_CIV
Descr: MITAC Group chairman
LOCATION: Type COUNTRY
NAME: U.S.
LOCATION: Type COUNTRY
NAME: Taiwan

MUC(K) - The message understanding conference (cont.d)

> *Corpus-based technology evaluations like MUC and ATIS are just one style of evaluation; ultimately it is critical to demonstrate that systems can support users doing real tasks.*

Hirschman (1988:303)

The IREX project

- IREX (Information Retrieval and Extraction Exercise)
- evaluation-based project for IR and IE in Japanese, 1988-1999

NE	Example
ORGANIZATION	The Diet, IREX Committee
PERSON	(Mr.)Obuchi, Wakanohana
LOCATION	Japan, Tokyo, Mt.Fuji,
ARTIFACT	Pentium Processor, Nobel Prize
DATE	September 2, 1999; Yesterday
TIME	11 PM, midnight
MONEY	100 yen, $12,345
PERCENT	10%, a half

The ACE program - Automatic content extraction

A follow-up of MUC

- "The ACE research objectives are [...] the detection and characterization of Entities, Relations and Events." (2002)
- A mention is a reference to an entity (name, nominal or pronominal)
- It ran from 2002 to 2008 "For each entity, the annotation records the type of the entity (PER, ORG, GPE, LOC, FAC, VEH, WEA), subtype, class, and all the textual mentions of that entity."

There are no confirmed suspects yet, but officials say several Middle East groups are expected to be investigated.

Extent	**Type**	**REF-ATR**
[no confirmed suspects]	NOM.PER.GROUP.NEG	E1-REF
[officials]	BAR.PER.GROUP.USP	E2-REF
[several Middle East groups]	NOM.ORG.NON.SPC	E3-REF
[Middle East]	NAMPRE.GPE.CLUSTER.SPC.GPE	E4-ATR

The ACE program (contd.)

The ACE program, however, attempts to take the task "off the page" in the sense that the research objectives are defined in terms of the target objects (i.e., the entities, the relations, and the events) rather than in terms of the words in the text. For example, the so-called "named entity" task, as defined in MUC, is to identify those words (on the page) that are names of entities. In ACE, on the other hand, the corresponding task is to identify the entity so named. (Doddington et al., LREC2004)

Distant reading for sets of documents: task in 2008

The purpose of the Cross-Document (XDOC) task is to globally coreference these 50 ACE entities, and all ACE relations which contain them, over the 400 document corpus. These 400 documents will have been previously ACE annotated, so all ACE entities and relations will already be coreferenced within the documents.

The CoNLL shared task

Named entities are phrases that contain the names of persons, organizations, locations, times and quantities. The shared task of CoNLL-2002 concerns language-independent named entity recognition. We will concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups.

- they introduced the BIO format
- CoNLL-2002 had Dutch and Spanish
- CoNLL-2003 had German and English (had columns for POS and syntactic chunk as well)

```
Wolff B-PER
, O
currently O
a O
journalist O
in O
Argentina B-LOC
, O
played O
with O
Del B-PER
Bosque I-PER
in O
the O
final O
years O
of O
the O
seventies O
in O
Real B-ORG
Madrid I-ORG
O
```

BIO and BILOU schemes

An interesting issue, raised by Ratinov & Roth (2009) in CoNLL, is the representation scheme:

BIO Beginning, Inside, and Outside

BILOU Beginning, Inside, Last, Outside and Unit-length

They claim that

1. choice of encoding scheme has a big impact on the system performance
2. the less used BILOU formalism significantly outperforms the widely adopted BIO tagging scheme

The TimeML venture

TimeML is a specification language for event and temporal expressions in text. Events are things that happen or occur, or states or circumstances.

> *Temporally grounded events are the very foundation from which we reason about how the world changes. [..] event recognition drives basic inferences from text.* (Pustejowsky et al., 2003)

- discontinuous MWE event sequences are marked on their heads
- aspectualizers, and light verbs are also marked as events
- specific rules for causation expressions for events
- generics not marked
- differences between DATEs, TIMEs, DURATIONs and SETs of times
- marking signals (temporal prepositions, conjunctions, and modifiers)

The Extended Named Entity Hierarchy

- The Extended Named Entity Hierarchy is designed and developed to meet increasing needs for wider range of NE types.
- For example, it separates between PERSON, GOD, and CHARACTER
- Each NE has a set of attributes. PERSON has 20, GOD has none, CHARACTER has 16

  (...) an artificial classification of names. Such a classification so heavily relies on an individual's subjective impression on names that the author attached his name to the definition.

Attributes (16)	Examples of Attribute Values
Type of Work	film, childrens' books, animated comedy
Country	United States, England, Japan
Type	duck, little girl, starfish
Name of Work	SpongeBob SquarePants, Journey to the West, The Wise Little Hen
Costar	Mickey Mouse, Olive Oyl, Zhu Bajie
Creator	Walt Disney, P. L. Travers, Elzie Crisler Segar
Place of Setting	London, Tokyo, underwater city of Bikini Bottom
Date of First Appearance	1871, 1929, 1934
characteristics	sarcastic, very unintelligent, optimistic
Appearance	with large blue eyes, overweight, wears a sailor shirt, cap, and a red or black bowtie
Number of Works	89, more than 120, 125
Author of Work	Bud Sagendorf, Hy Eisman, Tom Sims
Date of Movie	1931, 1939, 1964
Occupation of Creator	film producer, animator, cartoonist
Date of TV Program	1953, 1961, 1968
Date of Setting	14th century, 2020

HAREM

An evaluation contest for Portuguese, organized by Linguateca, with two main editions (2006 and 2008).

- devised for Portuguese proper nouns
- based on the way the language works, not the way the world is
- given that it was an evaluation contest to improve Portuguese processing, no matter the application, a lot of work was done for accounting for all sorts of application contexts: participants could address all subsets of categories
- special attention to vagueness, and to identification alternatives
- context-dependent classification
- 10 major categories and 41 types: `PESSOA`, `ORGANIZACAO`, `TEMPO` (time), `LOCAL`, `OBRA` (work), `ACONTECIMENTO` (event), `ABSTRACÇÃO` (`abstraction`), `COISA` (thing), `VALOR` (value), `VARIADO` (other)
- In the second edition, it had two extra subtasks: temporal identification TIME-ML-type and relation extraction among NEs.

Robinson Crusoe's *Friday*

Friday is difficult to parse, because weekdays are capitalized in English.
But *Sexta-feira* is obviously a proper name in Portuguese, because weekdays are not capitalized in Portuguese.

See also the example from Ratinov & Roth:

> *SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].*

Consequences of HAREM

- a lot of participants (8 systems, 14 runs; 5 systems, 20 runs; 10 systems, 27 runs): all in all, 17 systems
- golden data and evaluation service still available
- from the beginning, one of the complaints was that "it was not easy to compare to evaluations of other languages"
- spawned some, but not much, research on NER in Portuguese
- two books published, in Portuguese

TempEVAL

Evaluating Time Expressions, Events, and Temporal Relations according to TimeML.
Three stages

- TempEval(-1) - 2007 - temporal relations in English (event time, document creation time, and relations between main events in consecutive sentences. Events and times were given to the participants)
- TempEval-2 - 2010 - 6 languages, but participants only for English (16) and Spanish (2)
- TempEval-3 - 2012 - English and Spanish – full set of temporal relations, end-to-end task – events and times were not given to the participants.

SHINRA

Structured Knowledge, built on Wikipedia and Extended Named Entities

> *We believe that the structure of the knowledge should be defined top-down rather than bottom-up to create cleaner and more valuable knowledge bases. Instead of the existing, cumbersome Wikipedia categories, we should rely on a well-defined and fine-grained category definition.* (Sekine et al., 2020:177)

SHINRA2020-ML: a shared task at NCTIR 15 (2019-2020)

- 30 languages Wikipedia entities
- starting with the most 920K Japanese Wikipedia pages classified in ENE (the Extended Named Entity hierarcchy, with 220 categories)

Person	247,983	School	23,609
City	45,306	Literature	18,515
Artefact other	33,453	Movie	17,901
Broadcast Program	32,050	Train station	15,863
Company	26,746	Sports event	15,863

Table 1: Top 10 ENE categories in Japanese Wikipedia

What is there around of NER for literature?

There has been lot of work on (a subset of) NER in literary studies, mainly for characters and places.

- Character networks, and characters and places, have been dealt with e.g. for English (Lee & Yeung, 2012, Valaa et al., 2015, Dekker et al., 2019), Dutch (de Does et al., 2017), and German (Krug et al., 2018)
- Inferring latent character types, accounting for both referential and formal dimensions of narrative, called "personas", a distribution over 4 types of dependency relations (Bamman et al., 2014)
- it is often necessary to group different designations of the same character...
- do novels in an urban setting have more characters than those in a rural setting? (Elson et al., 2010)

Literary geography, or literary GIS

After identifying a place, one needs to geocode it.

- Comparing different subjective journeys (Cooper & Gregory, 2010)
- Using literature to map a city (Alves & Queiroz, 2013)

Just to give an idea of the problems that stand in the way of producing a map:

- *As heirs of D. Antónia Joaquina Xavier came three families from Lisbon, Evora and Tavira with the surname Nobre*
- *remembering the old journeys of his time, on horseback or litter, when in order to travel from Oporto to Lisbon one had to write a will*
- *Since Lisbon did not surrender, ...*
- *None of the travellers had received news from Lisbon*

Plus: Often you infer a place by landmarks, not by the explicit name. And places change with time.

In our COST action: distant reading

"Unusual" conditions: data

- literary texts: novels from 1840-1920
- old ortographies, OCR-errors
- many languages and linguacultures

Questions:

- which categories would make most sense for NER in literary texts?
- which categories/work can be done by off-the-shelf systems?
- how to harmonize different languages?

Three steps so far

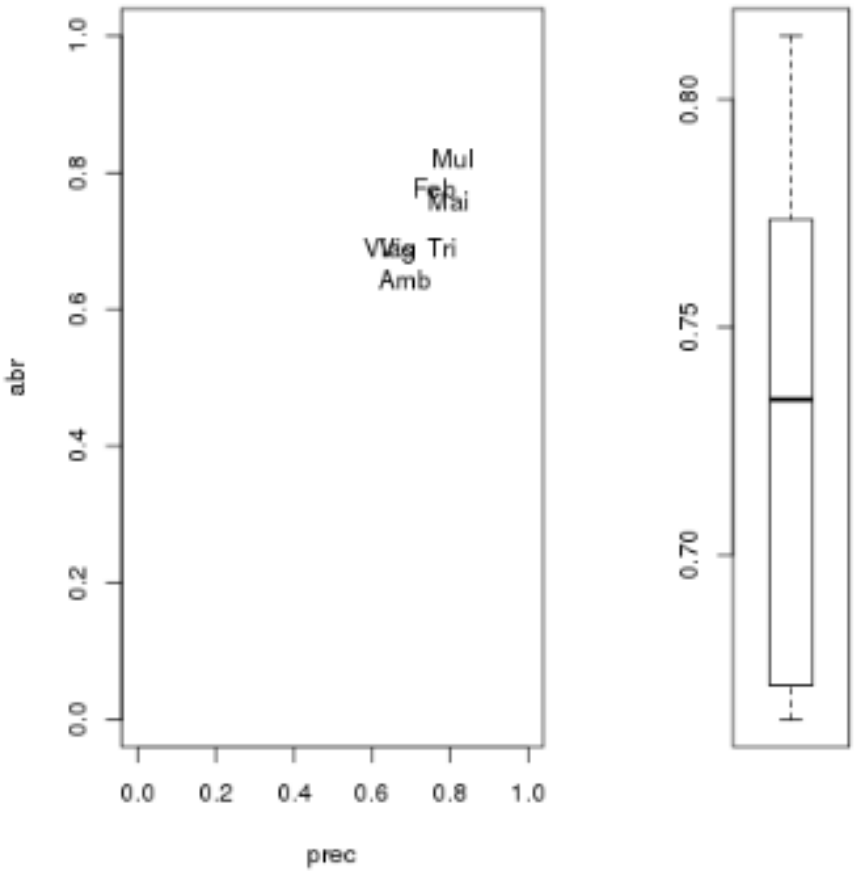
- Discussion with literary scholars (WG3) made us (WG2's NER-subgroup) decide on the following eight categories for all action languages: `Person`, `Role`, `Demonym`, `Location`, `Event`, `Work`, `Organization`, and `Other`.
- Thanks to several volunteers, we performed in 2018 an initial annotation of more than 50,000 words each in six languages (eng, fra, por, hun, nor, slv), extended to eight (cze, srb) in 2019, reported in Stanković et al. (2019), and to nine (deu) in 2020.
- For four languages (eng, fra, por, srb) we compared, based on the golden material, two automatic NER systems as far as PERS and LOC are concerned. See data in Frontini et al. (2020).

Currently, annotation for level 2 is taking place for several collections. However, not all languages will use the same categories and/or guidelines. Slovene has (so far?) only used Person, Organization, Location and Other.

Two experiments with Portuguese

1. comparing canonical works in modern ortography with non-canonical in old ortographies in the golden collection
2. analysing the performance of an automatic system on 8 complete books

Preliminary conclusions (Santos et al., 2020): one can go a long way by finely adjusting the lexicons.



Some important concepts

- One very important concept is what has been variously called grounding, or referencing of named entities, which is to bridge between a textual description, and an "outside" one (an encyclopedia entry, an URL, a Dewey code, geographical coordinates, etc.). A different knowledge representation system, in short.
- So, NER should be thought as textual/NLP task, which is then followed by a "referencing" task.
- Due to the unique characteristics of natural language, such as vagueness or underspecification, this referencing can be quite tricky.
- Also, this referencing may be time dependent in several ways.

Concluding remarks

The goal of this presentation was to survey the several different alleys "named entity recognition" has travelled around. There is not a single task definition and/or understanding of what it is all about.

- From general NLP/IR to specific task and back
- Bottom-up or top-down definition
- A model of the world, or of the language?
- Application-dependent, or general purpose?
- Format-dependence? Evaluation leeway?

Defining and operationalizing NER in our context

- For a collection such as ELTeC, a representation of (novels of) several literatures, we can define the task as identifying the characters, historical or fictional, the places, real or fictional, the events, historical or fictional, ...
- The challenge is to develop systems that do this to a satisfying extent, while remaining aware of differences between languages, and literatures – not forcing commonalities where they do not exist

References

A reference list is available from `https://github.com/distantreading/WG2/tree/master/NER_TS/References`.
Some relevant sites:

- `https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications`
- Language-Independent Named Entity Recognition at CoNLL
  - `https://www.clips.uantwerpen.be/conll2002/ner/`
  - `https://www.clips.uantwerpen.be/conll2003/ner/`