

Named Entities and Coreference. Annotation.

Rodrigo Agerri
IXA NLP Group
UPV/EHU
rodrigo.agerri@ehu.eus

Abstract

Looking at real examples to annotate and debate Named Entities and Coreference.

1 Introduction

Named Entity Recognition (NER) is one of the most commonly used low-level linguistic annotations and analysis for Natural Language Processing (NLP) tasks such as Information Retrieval (IR), Question Answering (QA) or Machine Translation (MT). Named Entities (NEs) are viewed as *rigid designators* that uniquely denote an entity (Kripke, 1980). These entities usually include proper names of locations such as Spain, France, Mediterranean Sea, English Channel, Alps, etc., organizations/institutions (Microsoft, IBM, BBC, European Central Bank) and people’s names (John McEnroe, King Albert II, Queen Elizabeth, Nixon).

Intuitively, NEs are seen as rigid designators because they uniquely refer to an entity in the world, as opposed to non-rigid designators (NRDE) which merely refer to an entity in a non-unique manner. Classic examples of non-rigid designators are those corresponding to descriptions (Russell, 1905) which are used to describe an entity in the world. For example, “the current British prime minister” is a description of a person who is (at the time that sentence was written) the prime minister of the UK. Nowadays the reference of this description is “David Cameron”, but that changes with the time at which the description is used, namely, during the second part of the Second World War the referent would have been “Winston Churchill”. Common non-rigid designators include “New York Times correspondent”, “the White House speaker”, “the president of the European Commission”, “IBM headquarters”, etc.

In order to perform NER as accurately as possible, it is crucial to decide, before doing any annotation, which types of entities are we annotating and to do so consistently. If we limit ourselves to rigid designators then we should consistently annotate “New York Times” instead of about a *journalist* of the New York Times; the same applies to “White House” and the *speaker* of the White House, “European Commission” and the *president* of the European Commission and “IBM” and the *headquarters* of IBM.

NER is traditionally taken as an important attribute to establish coreference, but so far the role that coreference could play to improve the NER task has been less researched. In any case, if we were to annotate a corpus with coreferences, this annotation should also be consistent with the NE annotation.

2 Ontonotes Example

```

-----
Plain sentence:
-----
The most important thing about Disney is that it is a global brand.

Treebanked sentence:
-----
[Zhou_liangshuyi] The most important thing about Disney is that it is a global
brand .

Tree:
-----
(TOP (S (CODE [Zhou_liangshuyi])
(NP-SBJ (NP (DT The)
(ADJP (RBS most)
(JJ important))
(NN thing))
(PP (IN about)
(NP (NNP Disney))))))
(VP (VBZ is)
(SBAR-PRD (IN that)
(S (NP-SBJ (PRP it))
(VP (VBZ is)
(NP-PRD (DT a)
(JJ global)
(NN brand)))))))
(. .)))

Leaves:
-----
0 [Zhou_liangshuyi]
   coref: IDENT          000-m_26 0-0   [Zhou_liangshuyi]
1 The
2 most
3 important
4 thing
5 about
6 Disney
   coref: IDENT          000-21 6-6   Disney
   name: ORG              6-6     Disney
7 is
   sense: be-v.2
   prop: be.01
   v          * -> 7:0 is
   ARG1       * -> 1:2 The most important thing about Disney
   ARG2       * -> 8:1 that it is a global brand
8 that
9 it
   coref: IDENT          000-21 9-9   it
10 is
   sense: be-v.1
   prop: be.01
   v          * -> 10:0 is
   ARG1       * -> 9:1 it
   ARG2       * -> 11:1 a global brand
11 a
12 global
13 brand
14 .
-----

```

3 Exercises

- (1) The disappearance of York University chef Claudia Lawrence is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father Peter Lawrence made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a 10,000 reward was offered to help find Miss Lawrence. Crimestoppers said the sum they were offering was significantly higher than usual because of public interest in the case.

(Example 1 of RTE-4)

- (2) American photojournalist James Nachtwey in a file photograph from May 18 2003 as he is awarded the Dan David prize in Tel Aviv for his outstanding contribution to photography. It was announced by Time magazine on Thursday, 11 December 2003 that Nachtwey was injured in Baghdad along with Time magazine senior correspondent Michael Weisskopf when a hand grenade was thrown into a Humvee they were traveling in with the US Army. Both journalists are reported in stable condition and are being evacuated to a US military hospital in Germany.

Caption 1470132 of the ImageCLEF-09 dataset (Paramita et al., 2009)

- (3) Nothing special really. Comfortable and clean but very boring decor in comparison to other NH hotels. I stayed in NH in Brussels and Zurich and I really liked them because of their modern and stylish design and big rooms. This one was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and 20 euros for breakfast!!! It was good but way overpriced! The best thing about the hotel was the location - city centre, 2min from a metro stop.

(review from OpenNER project)

References

- Kripke, S. (1980). *Naming and Necessity*. Harvard University Press.
- Paramita, M., Sanderson, M., and Clough, P. (2009). Diversity in photo retrieval: overview of the imageclefphoto task 2009. In *CLEF Working Notes 2009*.
- Russell, B. (1905). On denoting. *Mind*, 14:479–493.