Towards a Methodology for Named Entities Annotation

Karën Fort
INIST / LIPN
2 allée du Parc de Brabois,
54500 Vandoeuvre-lès-Nancy, France
karen.fort@inist.fr

Maud Ehrmann
XRCE
6 Chemin de Maupertuis,
38240 Meylan, France
Ehrmann@xrce.xerox.com

Adeline Nazarenko
LIPN, Université Paris 13 & CNRS
99 av. J.B. Clément,
93430 Villetaneuse, France
nazarenko@lipn.univ-paris13.fr

Abstract

Today, the named entity recognition task is considered as fundamental, but it involves some specific difficulties in terms of annotation. Those issues led us to ask the fundamental question of what the annotators should annotate and, even more important, for which purpose. We thus identify the applications using named entity recognition and, according to the real needs of those applications, we propose to semantically define the elements to annotate. Finally, we put forward a number of methodological recommendations to ensure a coherent and reliable annotation scheme.

1 Introduction

Named entity (NE) extraction appeared in the middle of the 1990s with the MUC conferences (*Message Understanding Conferences*). It has now become a successful Natural Language Processing (NLP) task that cannot be ignored. However, the underlying corpus annotation is still little studied. The issues at stake in manual annotation are crucial for system design, be it manual design, machine learning, training or evaluation. Manual annotations give a precise description of the expected results of the target system. Focusing on manual annotation issues led us to examine what named entities are and what they are used for.

2 Named Entities Annotation: practice and difficulties

Named entity recognition is a well-established task (Nadeau and Sekine, 2007). One can recall its evolution according to three main directions. The first corresponds to work in the "general" field,

with the continuation of the task defined by MUC for languages other than English, with a revised set of categories, mainly with journalistic corpora[1]. The second direction relates to work in "specialized" domains, with the recognition of entities in medicine, chemistry or microbiology, like gene and protein names in specialized literature[2]. The last direction, spanning the two previous ones, is disambiguation.

For each of those evaluation campaigns, corpora were built and annotated manually. They are generally used to develop automatic annotation tools. "To Develop" is to be understood in a broad sense: the goal is to describe what automatic systems should do, to help writing the symbolic rules they are based on, to learn those rules or decision criteria automatically, and, finally, to evaluate the results obtained by comparing them with a gold standard. The annotation process brings into play two actors, an annotator and a text. The text annotation must follow precise guidelines, satisfy quality criteria and support evaluation.

In the general field, the MUC, CoNLL and ACE evaluation campaigns seem to have paid attention to the process of manual NE annotation, with the definition of annotation guidelines and the calculation of inter-annotator (but not intra-annotator) agreement, using a back-and-forth process between annotating the corpus and defining the annotation guidelines. Nevertheless, some aspects of the annotation criteria remained problematic, caused mainly by "different interpretations of vague portions of the guidelines" (Sundheim, 1995). In the fields of biology and medicine, texts from specialized databases (PubMed and MedLine[3]) were annotated. Annotation guidelines

---

[1]See the evaluation campaigns MET, IREX, CoNNL, ACE, ESTER and HAREM (Ehrmann, 2008, pp. 19-21).

[2]See the evaluation campaigns BioCreAtIvE (Kim et al., 2004) and JNLPBA (Hirschman et al., 2005).

[3]www.ncbi.nlm.nih.gov/pubmed, http://medline.cos.com

were vague about the annotation of NEs [4], and few studies measured annotation quality. For the GE-NIA (Kim et al., 2003), PennBioIE (Kulick et al., 2004) or GENETAG (Tanabe et al., 2005) corpora, no inter- or intra-annotator agreement is reported. If NE annotation seems a well-established practice, it involves some difficulties.

As regards general language corpora, those difficulties are identified (Ehrmann, 2008). The first one relates to the choice of annotation categories and the determination of what they encompass. Indeed, beyond the "universal" triad defined by the MUC conferences (ENAMEX, NUMEX and TIMEX), the inventory of categories is difficult to stabilize. For ENAMEX, although it may be obvious that the name of an individual such as *Kofi Annan* is to be annotated using this category, what to do with *the Kennedys*, *Zorro*, *the Democrats* or *Santa Claus*? For the other categories, it is just as difficult to choose the granularity of the categories and to determine what they encompass. Another type of difficulty relates to the selection of the mentions to be annotated as well as the delimitation of NE boundaries. Let us consider the NE "Barack Obama" and the various lexemes that can refer to it: *Barack Obama*, *Mr Obama*, *the President of the United States*, *the new president*, *he*. Should we annotate proper nouns only, or also definite descriptions that identify this person, even pronouns which, contextually, could refer to this NE? And what to do with the various attributes that go with this NE (*Mr* and *president*)? Coordination and overlapping phenomena can also raise problems for the annotators. Finally, another difficulty results from phenomena of referential plurality, with homonyms NEs (*Java* place and *Java* language) and metonyms (*England* as a geographical place, a government or sport team).

Our experience in microbiology shows that these difficulties are even more acute in specialized language. We carried out an annotation experiment on an English corpus of PubMed notices. The main difficulty encountered related to the distinction required between proper and common nouns, the morphological boundary between the two being unclear in those fields where common nouns are often reclassified as "proper nouns", as is demonstrated by the presence of these names

in nomenclatures (*small, acid-soluble spore protein A* is an extreme case) or acronymisation phenomena (one finds for example *across the outer membrane (OM)*). In those cases, annotators were instructed to refer to official lists, such as Swiss-Prot[5], which requires a significant amount of time. Delimiting the boundaries of the elements to be annotated also raised many questions. One can thus choose to annotate *nifh messenger RNA* if it is considered that the mention of the state *messenger RNA* is part of the determination of the reference, or only *nifh*, if it is considered that the proper noun is enough to build the determination. Selecting semantic types was also a problem for the annotators, in particular for mobile genetic elements, like plasmids or transposons. Indeed, those were to be annotated in taxons but not in genes whereas they are chunks of DNA, therefore parts of genome. A particularly confusing directive for the annotators was to annotate the acronym *KGFR* as a proper noun and the developed form *keratinocyte growth Factor receptor* as a common noun. This kind of instruction is difficult to comprehend and should have been documented better.

These problems result in increased annotation costs, too long annotation guidelines and, above all, a lot of indecision for the annotators, which induces inconsistencies and lower-quality annotation. This led us to consider the issue of what the annotators must annotate (semantic foundations of NE) and, above all, why.

3 What to Annotate?

3.1 Various Defining Criteria

Ehrmann (2008) proposes a linguistic analysis of the notion of NE, which is presented as an NLP "creation". In the following paragraphs, we take up the distinction introduced in LDC (2004): NE are "mentions" refering to domain "entities", those mentions relate to different linguistic categories: proper nouns ("Rabelais"), but also pronouns ("he"), and in a broader sense, definite descriptions ("the father of Gargantua"). Several defining criteria for NE can be identified.

Referential Unicity One of the main characteristics of proper nouns is their referential behaviour: a proper noun refers to a unique referential entity, even if this unicity is contextual. We consider that this property is essential in the usage of NEs in NLP.

Referential Autonomy NEs are also autonomous from the referential point of view. It is obvious in the case of proper nouns, which are self-sufficient to identify the referent, at least in a given communication situation (*Eurotunnel*). The case of definite descriptions (*The Channel Tunnel operator*) is a bit different: they can be used to identify the referent thanks to external knowledge.

Denominational Stability Proper nouns are also stable denominations. Even if some variations may appear (*A. Merkel/Mrs Merkel*), they are more regular and less numerous than for other noun phrases[6].

Referential Relativity Interpretation is always carried out relatively to a domain model, that can be implicit in simple cases (for example, a country or a person) but has to be made explicit when the diversity in entities to consider increases.

3.2 Different Annotation Perspectives

The defining criteria do not play the same role in all applications. In some cases (indexing and knowledge integration), we focus on referential entities which are designated by stable and non-ambiguous descriptors. In those cases, the NEs to use are proper nouns or indexing NEs and they should be normalized to identify variations that can appear despite their referential stability. For this type of application, the main point is not to highlight all the mentions of an entity in a document, but to identify which document mentions which entity. Therefore, precision has to be favored over recall. On the other hand, in the tasks of information extraction and domain modelling, it is important to identify all the mentions, including definite descriptions (therefore, coreference relations between mentions that are not autonomous enough from a referential point of view are also important to identify).

As it is impossible to identify the mentions of all the referential entities, the domain model defines which entities are "of interest" and the boundary between what has to be annotated or not. For instance, when a human resources director is interested in the payroll in the organization, s/he thinks in terms of personnel categories and not in terms of the employees as individuals. This appears in the domain model: the different categories of persons (technicians, engineers, etc.) are

modelled as instances attached to the concept CAT-OF-EMPLOYEES and the individuals are not represented. On the opposite, when s/he deals with employees' paychecks and promotion, s/he is interested in individuals. In this case, the model should consider the persons as instances and the categories of personnel as concepts.

Domain modelling implies making explicit choices where texts can be fuzzy and mix points of view. It is therefore impossible to annotate the NEs of a text without refering to a model. In the case of the above experiment, as it is often the case, the model was simply described by a list of concepts: the annotators had to name genes and proteins, but also their families, compositions and components.

4 Annotation methodology

Annotation guidelines As the targeted annotation depends on what one wants to annotate and how it will be exploited, it is important to provide annotators with guidelines that explain what must be annotated rather than how it should be annotated. Very often, feasibility constraints overcome semantic criteria,[7] which confuses annotators. Besides, it is important to take into consideration the complexity of the annotation task, without excluding the dubious annotations or those which would be too difficult to reproduce automatically. On the contrary, one of the roles of manual annotation is to give a general idea of the task complexity. The annotators must have a clear view of the target application. This view must be based on an explicit reference model, as that of GENIA, with precise definitions and explicit modelling choices. Examples can be added for illustration but they should not replace the definition of the goal. It is important that annotators understand the underlying logic of annotation. It helps avoiding misunderstandings and giving them a sense of being involved and committed.

Annotation tools Although there exists many annotation tools, few are actually available, free, downloadable and usable. Among those tools are Callisto, MMAX2, Knowtator or Cadixe[8] which was used in the reported experiment. The features

---

[6]*A contrario*, this explains the importance of synonyms identification in domains where denominations are not stable (like, for instance, in genomics).

[7]"In [src homology 2 and 3], it seems excessive to require an NER program to recognize the entire fragment, however, 3 alone is not a valid gene name." (Tanabe et al., 2005).

[8]http://callisto.mitre.org, http://mmax2.sourceforge.net, http://knowtator.sourceforge.net, http://caderige.imag.fr

and the annotation language expressivity must be adapted to the targeted annotation task: is it sufficient to type the textual segments or should they also be related? is it possible/necessary to have concurrent or overlapping annotations? In our experiment on biology, for instance, although the annotators had the possibility to mention their uncertainty by adding an attribute to the annotations, they seldom did so, because it was not easy to do using the provided interface.

Annotation evaluation Gut and Bayerl (2004) distinguishes the inter-annotator agreement, which measures the annotation stability, and the intra-annotation agreement that gives an idea on how reproducible an annotation is. The inter- and intra-annotator agreements do not have to be measured on the whole corpus, but quite early in the annotation process, so that the annotation guidelines can be modified. Another way to evaluate annotation relies on annotator introspection. Annotators are asked to auto-evaluate the reliability of their annotations and their (un)certainty attributes can be used afterwards to evaluate the overall quality of the work. Since we did not have several annotators working independently on our biology corpus, we asked them to indicate the uncertainty of their annotations on a carefully selected sample corpus. 25 files were extracted out of the 499 texts of our corpus (5%). This evaluation required only few hours of work and it enabled to better qualify and quantity annotation confidence. The annotators declared that around 20% of the total number of annotation tags were "uncertain". We observed that more than 75% of these uncertain tags were associated to common nouns of type *bacteria* and that uncertainty was very often (77%) linked to the fact that distinguishing common and proper nouns was difficult.

More generally, a good annotation methodology consists in having several annotators working independently on the same sample corpus very early in the process. It allows to quickly identify the disagreement causes. If they can be solved, new recommendations are added to the annotation guidelines. If not, the annotation task might be simplified and the dubious cases eliminated.

5 Conclusion and Prospects

In the end, two main points must be considered for a rigorous and efficient NE annotation in corpus. First, as for the content, it is important to focus, not on *how* to annotate, but rather on *what* to annotate, according to the final application. Once specified what is to be annotated, one has to be cautious in terms of methodology and consider from the very beginning of the campaign, the evaluation of the produced annotation.

We intend to apply this methodology to other annotation campaigns of the project we participate in. As those campaigns cover terminology and semantic relations extraction, we will have to adapt our method to those applications.

References

Maud Ehrmann. 2008. *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Ph.D. thesis, Univ. Paris 7.

Ulrike Gut and Petra Saskia Bayerl. 2004. Measuring the reliability of manual annotations of speech corpora. In *Proc. of Speech Prosody*, pages 565–568, Nara, Japan.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(1).

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:180–182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proc. of JNLPBA COLING 2004 Workshop*, pages 70–75.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004 Workshop: Biolink*. ACL.

LDC. 2004. ACE (Automatic Content Extraction) english annotation guidelines for entities. Livrable version 5.6.1 2005.05.23, Linguistic Data Consortium.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigaciones*, 30(1):3–26.

B. Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proc. of the 6th Message Understanding Conference*. Morgan Kaufmann Publishers.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and John Wilbur1. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *Bioinformatics*, 6.