International Conference on Computational Semantics, Interoperable Semantic Annotation workshop (Oxford 2011)

A Deep Ontology for Named Entities

Gil Francopoulo
Tagmatica
126 rue de Picpus
75012 Paris France
gil.francopoulo@tagmatica.com

François Demay
(Independent Consulting)
26, rue de Rochefort
78120 Clairefontaine France
francois@edemay.com

Abstract

The number of documents published every day on the web becomes huge. Manual construction of metadata for each page being hard and unreliable, automatic semantic annotation seems to be the only practical solution to meet these increasing needs for information extraction and retrieval. In this respect, ontology design plays a key role: a common ontology permits a good interoperability between software components at the level of semantic annotations, but also allows the final user to easily interpret the result of the semantic extraction. In this paper, we present an operational, industrial and extended ontology for the processing of newspapers and blogs.

1 Introduction

Ontologies for named entities (NE) are designed to meet increasing needs for NE types. These ontologies originated from the set defined by MUC [Grishman 1996] with 7 categories (people, organization, location, time, date, money, percentage expressions). The number of types was increased to 93 (29 types and 64 subtypes) by BBN [Brunstein 2002]. The types were extended in several steps by Sekine to reach 150 and then 200 NE types [Sekine 2002, 2004][1]. For the history of NE typing, see [Ehrmann 2008].

2 Objectives

Just dealing with newspapers, newswires and blogs, the number of documents published every day on the web becomes huge. Only a subset of these contents is annotated with reliable metadata, and, of course, except for certain known sources, frequently, we don't know whether the metadata is correct or wrong[2]. Another problem is that the metadata tags are heterogeneous and, therefore difficult to compare. The most reliable material still remains the text content.

Our aim is to filter, dispatch and classify documents for professional watchers. Named entities are very good clues for this purpose. From the point of view of the user, very simple algorithms may be implemented: one of them may be for instance to consider that if, in a document is met the name of a soccer player or a soccer club, this document deals with soccer.

3 Situation

The problem we faced was two-fold. First, the 200 Sekine's types were not precise enough in the micro-domain where the system has a rich net of information and where the requirements from our professional users are rather high. Secondly, it was not possible to adopt an ontology not suited for NE like the hierarchy of types of WordNet. As discribed in section-10 (ambiguities), we need some specific intermediate types in order to retain a certain level of indetermination. For instance, a splitting at a higher level between abstract and concrete objects is not useful. On the contrary, a node merging the geographical and the political aspect of a city is useful because this node provides a geopolitical type when the geographical and political aspect cannot be distinguished.

Another problem to solve was more on the human side than the scientific one. In a professionally system, some way or another, the

[1] See also the beta English V7 released when we wrote this article on http://nlp.cs.nyu.edu/ene

[2] For instance, very basic tags like language names are sometimes wrong. Typically, a page is written in one language, the content is translated and the original metadata is not modified accordingly.

type is presented to the user together with the ontology. That means that the user must understand the meaning of the type. The situation is different from the one in the 90's where an Academic player could develop a system in a laboratory and then propose the resulting system to professional users. Now watchers have ontologies, best practices habits, professional associations like IPTC[3] that recommend lists of codes. This new context has an impact on the methodology: the current users are rather mature and want to be associated with the design.

This is not to say that the situation is idyllic. Some users find that the professional ontologies cover the majority of their needs but some others think that the recommended lists and structures are too complex. We tried to accommodate with this situation.

Our strategy was to build the ontology by hand from existing lists, with a co-operation between experts of encyclopedia writing, NLP practitioners and professional watchers. The role of an expert in encyclopedia writing is to propose an organization of types. The role of a NLP practitioner is to test against real texts to verify how the type matches against named entities. The role of the professional watcher is to verify whether the result is understandable.

4 Method for building the ontology

The main burden is to propose an organization of types.

The first step was to split the domain according to Sekine's high level types. This is not to say that we did not modify afterwards these assumptions, but we took these types as a starting point. These modifications have been applied based on experiments on unknown named entities as explained in the following section "The ontology".

Starting from a type, a series of steps were tempted.

Step#1: is there a sub-tree which may be taken from a recommended professional list?

Step#2: is there a sub-tree in an encyclopedia? Usually, a kind of classification of the articles has been implemented in order to help the managers / editors master a certain balance between subjects or a relative completeness. These classifications were always some kind of topic tree where each level of the hierarchy could mix different nature of information: nature of the item being written about, place, time etc. To have a much better way to master the content, this type of classification can be replaced and complemented by a more systematic categorization scheme where each article would be located in a multidimensional space: each coordinate of this space would correspond to the answer (for the item) to the W questions : what / who is it?, where is it located in a subject / knowledge cartography?, where is it located in the (present or a previous) geopolitical area?, when is the item taking place? When this kind of categorization is implanted it becomes very easy to make a thorough analysis of the named entities, for proper names, strictly speaking, at least.

Step#3: Another way, when no such possibility exists as explained above, one can use the first sentence of the article of the encyclopedia: all the most relevant features defining the item are there, and one can retrieve a lot of metadata leading usually to the named entities characterization.

Step#4: Is there is any fact box (infobox in Wikipedia's terminology) that may also be used?

Step#5: Last, when categories have been given to the article, some of these may also be used. In Wikipedias, this has been done. Unfortunately, there has been usually no consistence in the rules for the categorization (and their implementation) even for articles in one and the same language and much more less consistency between categories in different language contents.

One of the author of the current work has been able to implement the systematic categorization (as described above with multidimensional localization) to a certain number of big encyclopedic contents in French, English and German for the Encyclopedia Universalis (15 years, since 1965), Larousse (20 years), Encarta (3 years) and then for a Chinese-French lexicon. This experience of 40 years has been very useful in building by hand the current ontology.

[3] www.iptc.org

5 Design principles

The Tagmatica's deep ontology for named entities is designed mainly for information extraction from newspapers, newswires and blogs. The type names are labeled in English and the meaning of each type is the definition in English but we took great care to respect a language neutrality because the NE extraction is not restricted to English. Currently, the same ontology is used for three languages: French, English and Spanish. We didn't have the need to create any specific sub-tree for a given language. Three domains are addressed on a fine grain basis: politics, economics and sports. These domains are rather general and universal. In other terms, we do not address technical domains like genomics or mechanics.

We distinguish the notion of type from the notion of role. A type (and sub-type) is considered as a rigid subdivision, in the sense that this is the reason why an entity is known. For instance, "Jacques Chirac" is considered as a politician, with the convention that we are dealing with the famous human being and not with an unknown person whose name is "Jacques Chirac", i.e. a homonym. This labeling is considered as type labeling because this is the reason why we know his name. We don't consider the fact that when he was a child (or a baby) he was not a politician. On the contrary, his role as president is not considered as a type. This information is managed at the instance level as a function name together with starting and ending dates.

6 The ontology

Our ontology is designed having in mind the named entity recognition (NER). This process faces two different situations: i) the name (or fragments of the name) is (or are) already recorded in the system and successfully recognized, ii) the name is unknown, which means that the immediate context must be interpreted in order to determine a type for the name.

Let's take two examples:

1) "Messi is supported in this by FC Barcelona, ...". With the convention that "Lionel Messi" is recorded in the database, the NER recognizes and determines that "Messi" is the name of a soccer player. Thus, a great level of detail can be determined, including the variants like "Leo Messi" (his usual name), "Lionel Messi" or "Lionel Andrés Messi" (his official full name) together with an URI to a Wikipedia entry for additional documentation.

2) In another situation, with the convention that "Marcel Dujardin" is an unknown person, the sentence "Marcel Dujardin drove too fast ..." cannot give such a level of detail. But, provided that "Marcel" is recognized as a male given name and the first letter in "Dujardin" is an uppercase one, the NER is able to determine that "Marcel Dujardin" is a pair of words combining a given name and a family name, i.e. a person name. But, aside from the sexual genre, we cannot determine any information about the usual activity of this person.

Based on the requirement that we must deal with both known and unknown names, we decided to design a deep ontology for types dedicated to known names and a level-1 types for unknown names. We structured the first level to determine a type in case of an unknown name. We may add, that in some cases, it is possible to determine a slightly more precise typing than just the first level, but the precision does not go very far. For instance, a flight identification code is recognized as a pair of specific uppercase letters (to be taken in a pick-list) and three digits. In other terms, the system is always able to determine a main type (the first level) and optionally, in the most favorable situation, a sub-type can be determined.

We defined and developed an ontology of 995 types based on Sekine's hierarchy, IPTC event types [EventsML-G2], geonames[4] and previous works in encyclopedia structuring.

7 First level types

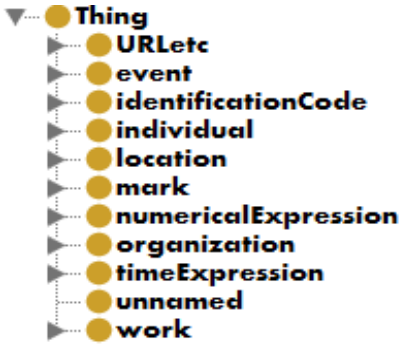
The level-1 has 11 types, as follows:

- **URLetc**, for filenames and URL,
- **event**, for event names like "Tour de France". This sub-tree is taken from the IPTC's registry for the types of events. This thematic classification is rather usual in the domain of newswires and professional watchers.
- **identificationCode** for all alphanumerical codes like flight number (e.g. AF 447) or ISBNs.
- **individual**, for an individual person. The person may be living (or has lived) or imaginary.

[4] www.geonames.org

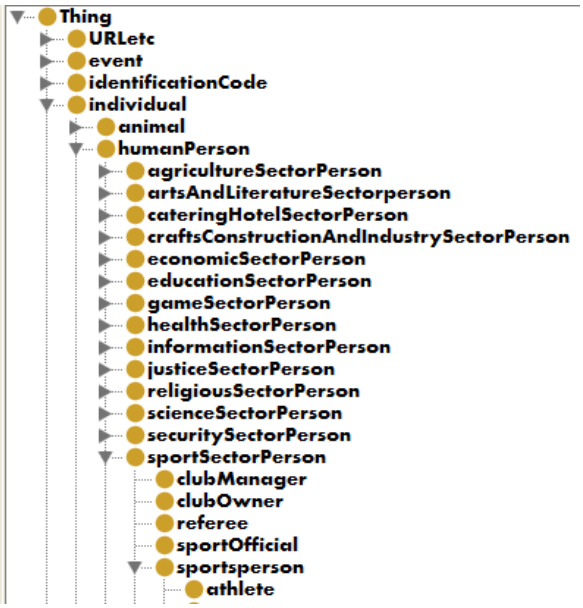
It is in general the name of a human being but the type may be used also for a pet or a plant. This type is not to be used for a group of people, see the organization item for this purpose.
- **location**, for geopolitical, geological and geographical entities. Continents and planets are also covered by this type.
- **mark**, for mark names like commercial trademarks, formats and protocols. Let's note that when the name is both a mark and the name of an organization, we adopt the convention that the name should be labeled as an organization.
- **numericalExpression**, for all forms containing a number with or without a unit. Examples are measures and percentages. This type is equivalent to NUMEX in MUC.
- **organization**, for a company, institution or a group of people.
- **timeExpression**, for a reference of time like dates, times, combinations of date and time. This type is equivalent to TIMEX in MUC.
- **unnamed**, for all the common nouns that are not in the other types and that are used as head of a noun phrase in the corpus. The main objective being to label named entities, this type cannot be used to directly mark a named entity. By means of the coreference, such a noun may be used to indirectly designate a named entity. For instance, in the text "Chirac ... The president ...", the function name "president" will be marked as unnamed and the coreference resolution module will link "Chirac" and "president"[5].
- **work**, for names of human works like movies, books, sculptures, songs etc.

To summarize, the first level is as follows in Protégé[6]:



8 One example

It is not possible to present in detail all sub-trees, so we will present only one example with a five-level depth for athlete:



9 Other levels

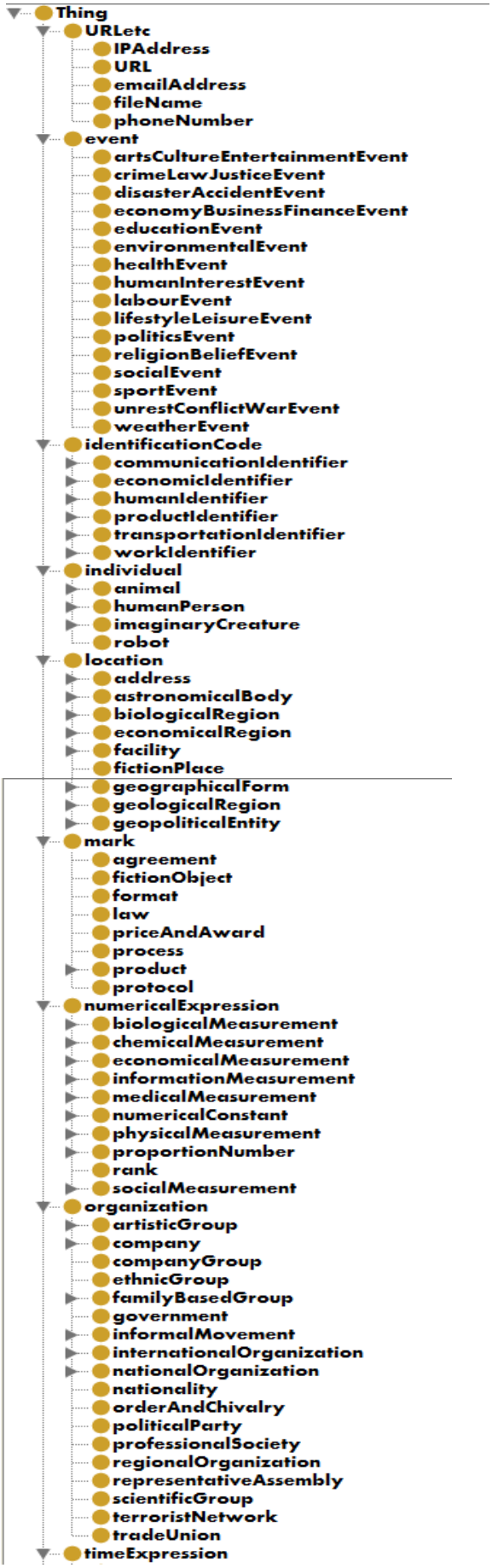
Depending on the types, the deepness of the hierarchy is between two and five. The ontology being rather large, we advice the interested reader to directly download the OWL file and use Protégé to browse through the different levels. The ontology is freely available on the Tagmatica's website[7].

The level-2 is as follows:

[5] As the notion of named entity has been extended from proper name to time expression in the 90's, we extend the notion of named entity to "non-proper" references to a named entity.

[6] http://protege.stanford.edu

[7] http://tagmatica.fr/doc/ontology.owl



10 Ambiguities

As mentioned in [Sekine 2002], "Japan" is normally used in a geographical sense, but sometimes it refers to the government of Japan (organization), as in "Japan announced a tax cut". This problem is a rather general problem that does not concern only the country names but concerns also all cities and villages. The NER is not able to distinguish the two senses on a reliable manner. So, we consider "Japan" as geographical and political entity (the GPE in ACE definition).

We adopt the same strategy for entities like airports that may be considered as geographical or as organizational entities. All these items are recorded under the node "facility". Instead of focusing on where and what the entity is, we prefer the usage aspect: "what is it for?".

11 The instances

The ontology is a hierarchy of types which is built by hand and is rather stable. On the contrary, the proper names (i.e. the instances), are automatically collected from different Wikipedia dumps[8] and may change frequently because new proper names appeared every day. For this purpose, a series of filters have been coded in Java in order to associate the types with the field names of Wikipedia's infoboxes. Most of the time, an instance is associated with only one type, but they are some exceptions like a famous judoka who is also deputy. We collect three Wikipedias (in French, English and Spanish), and from each file, we extract a selection of proper names. Then, the three results are merged together and finally merged with the current database. At present, we collect these names when needed but, in the near future, we plan to collect fresh data every week-end, systematically in order to be synchronous with new names. Let's add that a certain number of locations have been extracted from the

[8] http://download.wikipedia.org/backup-index.html

Geonames site but this work is not finished. At the moment, the total number of instances is 200 000.

This requirement to stick to the current state of the art of all fresh data published on the web has a certain number of constraints concerning the choice of the sources for updating the instances. Let's recall that our main application is newspapers processing. Wikipedia dumps are updated every four days (in average and when everything goes fine) without a precise date for each dump. Our instances will be updated every week-end, so the data are synchronous enough[9]. On the contrary, we cannot download DBpedia (see http://dbpedia.org) and the web of data which is computed from DBpedia (see http://linkeddata.org) because sources like DBpedia are updated every six months. This is for us a too long time. For the same reason, this requirement prevents us to use gazetteers included within frameworks like Gate (see http://gate.ac.uk). The update frequency is too low.

12 NE extraction

The NE extraction is not a stand-alone software module. It's a component of an hybrid industrial parsing scheme combining Hidden Markov Model implementations [Bikel 1997] and hand-written rules. The system been described elsewhere [Francopoulo 2008], we are not going to present the modules in detail. We may just add that the main modules are based on active learning techniques and that the whole system is a pipeline of modules for language detection, error recovering, chunking, syntactic parsing, coreference resolution and quotation extraction in a robust manner. The ontology of types and the instances are shared by the three processed languages. Each language is described in a specific lexicon called TagDico that conforms to the ISO standard ISO-24613 for NLP lexicons: LMF (for Lexical Markup Framework) [Francopoulo 2006].

13 Relation with standards

The NE extraction is consistent with the ISO Preliminary Work Item for the representation of named entities: ISO 24617-3 where the entities are annotated in a stand-off scheme in the spirit of the Linguistic Annotation Framework (LAF, i.e. ISO-24612) [Ide 2004].

[9] If we discover that it is not the case, we could refresh every day: the process is fully automatic.

With this respect, the labeling of NE is more powerful than inline annotation for difficult annotations in which the elements are not contiguous like "Bill and Hillary Clinton" and where the NER must detect two named entities with a distribution of the family name to the two given names. More traditional systems like BBN Named Entity Annotation (see www.anc.org/annotations.html#bbnne, for instance) cannot deal with such annotations because they are inline based. It should be noted that most systems do not deal with these problems (see [Erhmann 2008] for a discussion). The objective being to build an index, the two named entities must be recognized by the system.

Another difficult problem arises when one entity of a certain type is a sub-part of another entity with a different type. For instance, in: "the city of Michelin ..." where "city of Michelin" is a geopolitical entity (as a city) but where "Michelin" is the name of an organization. Let's note, that if an inline annotation scheme is used but with the option of embedding different levels of annotation, the annotation is possible, on the contrary of the first example where there is no way to address the problem. Again, the objective being to build an index, the two named entities must be recognized by the system.

Concerning XML serialization, the physical file is coded in OWL as defined by W3C at www.w3.org/2004/OWL.

14 Evaluation

There is no quantitative evaluation. Evaluation is important, but we have no budget for this task and this is not the right period. Our users make comments and the system is modified almost every day.

15 Conclusion

Following rather practical lines of action and after some long talks and negotiations based on an extended experience of ontology structuring, we created a deep ontology of types for named entities representation and automatic recognition with a fine set of interoperable semantic annotations.

The domains we currently address being politics, economics and sports, we targeted rather general domains. We don't claim that our

ontology could be easily extended to include deep technical domains like genomics or mechanics. At first view, these domains require separate and specific ontologies. But what is possible, is to extend selected sub-trees to cover more deeply a specific application domain. In the past, we already extended successfully some sub-trees for specific needs like airline, soccer or athletics domains, without any problem.

Our ontology begins to be effectively used and some sub-parts may need to be tuned or extended in the near future based on the user' feedbacks. We welcome all comments and useful suggestions.

Acknowledgements

The work presented here was partially funded by the System@tic competitiveness cluster http://www.systematic-paris-region.org within the Scribo project, see www.scribo.ws.

Reference

Bikel D., Miller S., Schwartz R., Weischedel R. 1997 "Nymble: a High-Performance Learning Name Finder" ANLP-1997.

Brunstein Ada 2002 "Annotation guidelines for answer types", BBN technologies report at www.ldc.upenn.edu/Catalog/docs/LDC2005T33 (on Dec 2010)

Ehrmann M. 2008 "Les entités nommées, de la linguistique au TAL", PhD thesis, Univ. Paris 7.

EventsML-G2 International Press Telecommunications Council (IPTC) www.iptc.org

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF), LREC-2006, Genoa.

Francopoulo G. 2008 TagParser: well on the way to ISO-TC37 conformance. ICGL-2008, Hong Kong.

Grishman R., Sundheim B 1996 "Message Understanding Conference - 6: A Brief History", COLING-96.

Ide N., Romary L. 2004 International Standard for a Linguistic Annotation Framework, Journal of Natural Language Engineering, 10:3-4, 211-225.

Sekine S., Sudo K., Nobata C. 2002 Extended Named Entity Hierarchy, LREC-2002, Las Palmas.

Sekine S., Nobata C. 2004 Definition, dictionaries and tagger for Extended Named Entity Hierarchy, LREC-2004, Lisbon.