Reference number of working document: **ISO/TC 37/SC 4 N605**

rev05

Date: 2010-12-19

ISO PWI 24167-3

Committee identification: ISO/TC 37/SC 4/WG 2

Secretariat: KATS

Language resource management - Semantic annotation framework - Part 3: Named entities

Gestion de ressources linguistiques - Cadre d'annotation sémantique - Partie 3: Entités nommées

Document type: International standard
Document subtype: if applicable
Document stage: XX.XX
Document language: en

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

International Standard 24617-3 was prepared by Technical Committee ISO/TC 37, *Terminology and other language resources*, Subcommittee SC 4, *Language resource management*.

This file has been edited using open source tools: StarUML 5 for the diagrams and OpenOffice 3.2 for the text.

Introduction

The aim of this specification is to propose a consensual annotation scheme for Named Entities (NEs).

The current specification is developed under the aegis of the ISO Semantic Annotation Framework (SemAF) where it is named SemAF-NE.

Named entities are very popular in the tasks of information extraction processing because they are often ideal record values [1].

The main areas of application are:

- Question answering: to determine for instance where a given proper name is mentioned within a corpus;
- Automatic, or semi-automatic construction of ontologies;
- Thematic computation: to determine what is the central theme of a given text within a corpus;
- Comparison: to determine the degree of similarity of different documents within a corpus;
- Machine translation to look up NEs into a special lexicon because usually they have a special status concerning translation;
- Message identification for automatic filtering, classification, and dispatching.

All these applications share a common aim that is to improve information management by processing the content of the documents, notably in the context of the Semantic web.

The specification will be used in two different situations:

- in annotations where the NEs are statically recorded in a resource, for instance, as a source for machine learning techniques;
- as a dynamic structure produced by an automatic system.

The current specification makes the assumption that these two situations (dynamic and static) share the same data structure model.

The objectives of this specification are to:

- allow the comparison and merging of different pre-existing annotations;
- provide a “best practice” based specification for new annotations that are thus natively interoperable with each other and with pre-existing annotations;
- allow NEs to be integrated into other annotation schemes like TimeML and ISO-Space;
- and ultimately to allow software developers to provide common tools.

The associated working group is made of:
Gil Francopoulo (France), as project leader,
Alex Fang (People's Republic of China),
James Pustejovsky (USA),
Kiyong Lee (South Korea),
Harry Bunt (Netherlands),
Thierry Declerck (Germany),
Antonio Toral (Italy),
Koiti Hasida (Japan),
Kais Haddar (Tunisia).

1. Scope

The identification of an NE comprises the different elements that describe the NE, including:

- The semantic type of an NE. Examples of type are “location“ and “organization“.
- The source type that describes how the NE was recognized. Examples are “lexicon“, “associated with an introducer“, “pattern based“ etc.
- The kind of word form used in the occurrence of the NE in the text. Examples are “abbreviation” and “full form”.
- The structure that decomposes the NE into substructures like, for instance, given name and family name.

It should be noted that the mechanism that deals with co-reference is not within the scope of the current specification, even if such a mechanism can refer to one or several NEs. Notably, the annotation of the variants is also not addressed where the challenge is to link the named entity occurrence "Jacques Chirac" with another occurrence like "J. Chirac". The co-reference from a pronoun to a named entity is not addressed.

2. Normative References

The following normative documents contain provisions that, through reference in this text, constitute provisions of ISO 24617-3. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on ISO 24613 are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

ISO 12620:2009 Computer applications in terminology – Data categories – Data category registry
ISO DIS 24611 Language resource management – Morphosyntactic annotation framework (MAF)
ISO DIS 24612 Language resource management – Linguistic annotation framework (LAF)
ISO 24615:2010 Language resource management – Syntactic annotation framework (SynAF)

3. Terms and Definitions

For the purposes of this International Standard, the following terms and definitions apply:

Chunk

flat sequence of words which contains more than one word and which does not contain any sub-chunk

Clause

unit of grammatical organization smaller or equal to the sentence but larger than phrases and words

Note: the traditional classification is of clausal units into main (independent or superordinate) and subordinate (or dependent) clauses.

Example: the boy arrived (main clause) before she came (subordinate clause)

Main clause

clause which is not subordinated to another clause

Named entity
NE

element made of one or several word forms which is classified into a predefined category and which has one or several referents that may be identified by a program or a human being

Examples: Bonn (classified as a location), Vanessa Paradis (classified as a person name)

Named entity recognition
NER
Entity identification

subtask of information extraction that seeks to locate NE in texts and classify these NEs into predefined categories

Note: some systems are hand-written rule based (i.e. symbolic), some systems are annotated corpus based (i.e. statistical) and other systems are hybrid ones, combining symbolic and statistical techniques

Phrase

structural element built around a main word (when there is one), formed of zero, one or more words and lacking the subject-predicate structure typical of clauses

Note: a phrase may embed sub-phrases. Several types are usually distinguished, e.g. noun phrase, adverb phrase, preposition phrase, verb phrase.

Example: the boy (noun phrase)

Prepositional chunk

chunk beginning with a preposition

Word form

contiguous or non-contiguous fragment from a speech or text sequence identified as an autonomous lexical item (from ISO-24615).

4. Key standards

Unicode

SemAF-NE is Unicode compliant and presumes that all data are represented using Unicode character encodings.

ISO 12620 Data Category Registry (DCR)

The designers of SemAF-NE conformant data shall use data categories from the ISO 12620 Data Category Registry (DCR) [3].

Unified Modeling Language (UML)

SemAF-NE complies with the specifications and modeling principles of UML as defined by the Object Management Group (OMG) [2]. This current specification uses a subset of UML that is relevant for the needed linguistic description.

5. Linguistic and processing requirements

Semantic type

Every NE may be labelled by a semantic type like “organization” or “individual”. It is rather difficult to freeze a rigid pick-list of tags because different developers use different lists of tags. Some systems use less than ten tags (for instance for MUC evaluation [5]) while some other systems use more than one hundred different tags, see for instance Sekine’s [7][8] or IPTC’s [4] long list of tags. Some systems use a thousand of types organized hierarchically, see for instance Tagmatica's ontology [9]. Some systems consider dates as NEs while some other systems do not. Some systems use exclusive values like “location” and “political entity” while other systems allow polysemic values like GPE (for geopolitical entity) as in the ACE evaluation [6].

The level of detail is a key point in this difference but the topic of the list is also important: some set of tags are rather general while some lists are specific to biomedicine or politics. Some lists of tags are a flat list while some other tags are hierarchically organized in an ontology or in a user reserved vocabulary. Some systems use two levels of typing: one for a top level type (called a type) and one for a more detailed type (called a subtype), see, for instance, the 29 types and 64 sub-types used by the BBN's system [10].

The most famous are the three values called ENAMEX in the foundation work of MUC, where ENAMEX was the name for organization, individual and location. Currently, most general systems use more or less the same seven or eight top level types but their employment is not uniform.

Kind of graphical form

Every NE may be labelled by a morphosyntactic attribute that describes the sort of fragment of text from a graphical point of view.

Example: “abbreviation” and “full form” that are to be taken in the morphosyntactic profile of the ISO data category registry.

Mono vs multi-words aspect and syntactic constituency

Basically, and with respect to the character strings of the words in the sentence, an NE is a fragment of text.
An NE may comprise only one word, like “Brad” in “Brad is late”.
An NE may comprise more than one word like “Brad Pitt” in “Brad Pitt is late”.
An NE may not match a full syntactic chunk like “I talked to Brad Pitt” where the prepositional chunk is “to Brad Pitt” and the NE is “Brad Pitt”. In other terms, the NE is smaller than the chunk because the preposition is not a part of the NE.
An NE may span several chunks like in the NE “King of Belgium” where “of Belgium” is the modifier of “King”.

Continuity

The fragment of text is usually a continuous sequence of words like "Los Angeles" in "Los Angeles is a city", but in some special cases, the NE may be discontinuous like in "Bill and Hillary Clinton came this morning" where "Bill Clinton" is a discontinuous NE because the family name of "Bill" is factorized after "Hillary". Let us insist on the fact that it is important to compute two different NEs: one for “Bill Clinton” and one for “Hillary Clinton”. Producing “Bill” and “Hillary Clinton” is not a good option because "Bill" is not associated with the family name.

Multiple overlapping annotations

An NE may be annotated by different types on some shared spans of text. For instance, and supposing that the task is to extract organization and location on a sentence like “He works for IBM Korea”, an NER may produce two NEs: one for “IBM Korea” (labelled as organization) and one for “Korea” (labelled as location). Another example is "the city of Kodak" where "city of Kodak" is a location and "Kodak" an organization.

Structure

An NE may comprise substructures, for instance, a substructure for the given name and another substructure for the family name. Each substructure may be labelled specifically. For instance, in “John Smith” where it is assumed that “John” is the given name and “Smith” is the family name. Each substructure may comprise an unlimited number of words as in “José Sánchez de la Vega” where “Sánchez de la Vega” is the family name. Other examples of substructures for person names are titles or second given names.
The diversity of substructures may be rather large depending on different criteria like the type of application or the level of detail. Thus, the substructures for an application in biomedicine will not be the same as the ones for an application for newspaper processing.
The substructure may hold the result of a computation in order to normalize all the sub-parts of the named entity. For instance, dates may be normalized into predetermined sub-fields like day, month and year.

Named entity recognition

Different sorts of processes may be involved: an NE may be a sequence of words that is recognized because the words are in a predefined pick-list, e.g. "Peugeot", in "Peugeot makes cars" (assuming that “Peugeot” is in the lexicon). But in other situations, the sequence of words is a specific NE only because it respects a given pattern like, for instance, in "the Roche laboratory makes these pills", where “laboratory” is an introducer that functions as a trigger word and “Roche” is unknown before parsing. Obviously hybrid situations are possible for complex NEs where some parts are already recorded in a lexicon and some other parts are unknown but are located in a specific position or respect a given pattern.

In certain circumstances, it is rather difficult for an automatic process to determine a fully detailed structure because the structure depends on the semantic type and the type is too fuzzy. Examples are derived products like perfumes that are labelled according to fashion marks. For instance, in the sentence “I like Armani”. It is rather difficult for a program (and for a human being) to distinguish between the name of the perfume, the name of the clothes, the name of the company, or the name of the individual.

Languages

The notion of NE seems to be a notion that is common to a lot of languages, see, for instance, NE identification for 12 languages [11]. The current specification is not designed for a merely specific language or a specific family of languages.

Stand-off vs in-line annotation

The Linguistic Annotation Framework recommends the use of stand-off annotation, i.e. the construction of annotations in documents independent from the one containing the primary language data. Stand-off annotations refer to specific locations in the primary data by addressing character offsets or linguistic elements such as words, to which the annotation applies. Compared to in-line annotation, stand-off annotation has the advantages of respecting the integrity of the primary data and of allowing multiple annotations to be layered over a given primary document. For named entity annotation, in-line annotation would moreover be inadequate since some fragments of text can be discontinuous as in the example “Bill and Hillary Clinton”, mentioned before.

Bottom up transfer of values

Some values may be picked from the original material and transferred to the internal structure of an NE. For instance, the gender tag for a first name that comes from the morphosyntactic annotation level may be transferred to the whole NE. So, assuming “Deschaumelle” is an unknown proper noun but “Robert” is known as a masculine given name, a system may infer that “Robert Deschaumelle” refers to a masculine individual.
The annotation scheme should allow a place to store these values.

6. Modeling

Purpose

The purpose of this specification is to propose an annotation scheme for named entities.

Identification

The intrinsic identification of an NE comprises the different elements that describe the NE, including:

- The semantic type of an NE, with possibly a subtype;
- The kind of graphical form that describes the graphical form occurring in the texts;
- The source that describes how the NE was recognized;
- The morphosyntactic features based on MAF tagsets;
- The structure that optionally decomposes the NE into substructures.

Only the semantic type is mandatory. All other attributes are optional.

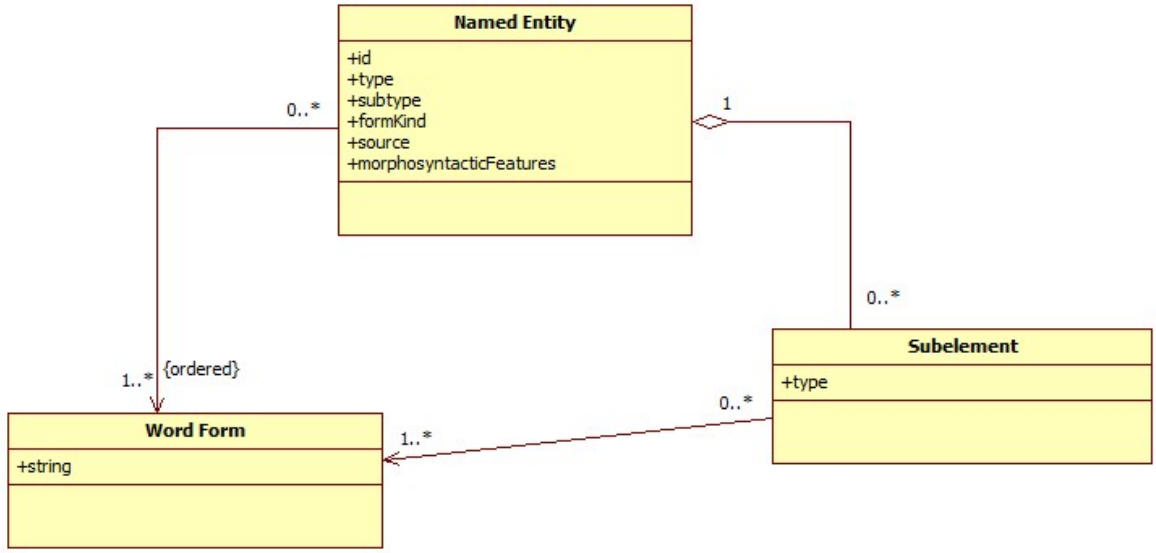
Model

The current specification does not determine a list of values for the semantic type, subtype, form kind or source. They are references to the values that are maintained in the ISO data category register.

The internal structure of an NE is an optional mechanism.

working notes:
* we are modeling the occurrence of a named entity. This occurrence is located in a text. This is not the description of a specific named entity that may be recorded in a lexicon.
* may be "source" is not a good name because it may be confused with the source of information (i.e. where the document comes from, like Reuters).
* on purpose, the terms "first name" and "last names" are avoided because in some languages like Japanese or Vietnamese, the given name and family name are usually written in the reverse order.
* question: do we allow recursive subelements?

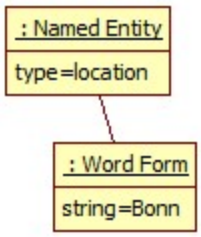
The model is specified by means of the following UML diagram:



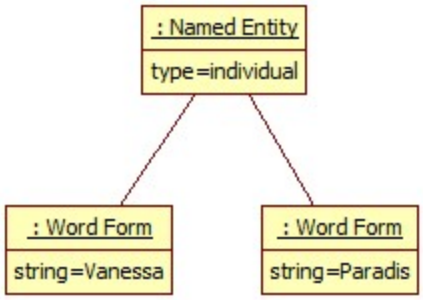
Annex-A (informative) Examples

Conventions
=> suggestion from James : give a small list of semantic types with a short definition otherwise the specification will be hard to understand

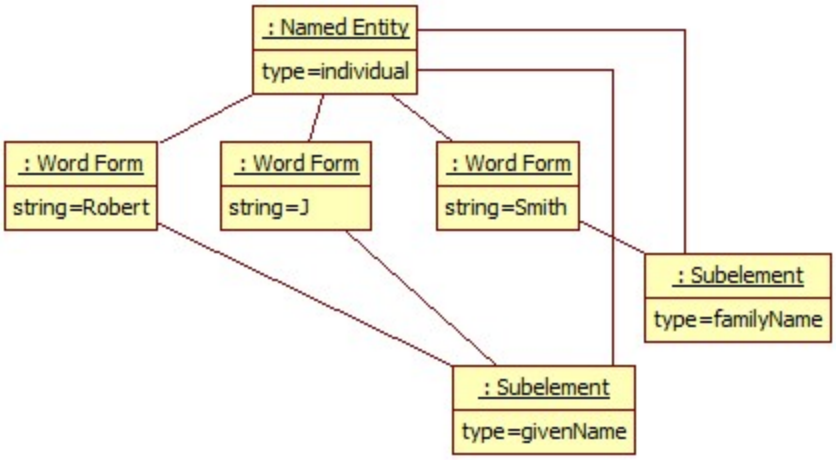
Simple examples
=> one naive example with only one wordform like Bonn



=> one naive example without any sub-structure like Vanessa Paradis



More complex examples
=> an example with a substructure for a person name (like Robert J. Smith) and recall that the substructure is optional.



=> example with a modifier (Smith Jr.)
=> an example with a substructure for a company name with an introducer (like Roche laboratory)
=> an example with a date (like "5th July" with a distinction between day and month)
=> a complex example with coordination (like Bill and Hillary Clinton)
=> a complex example with nested ENs (like "city of Michelin")

Annex-B (informative) XML DTD

Introduction

The following material is provided for information only. XML elements in the Document Type Definition (DTD) are transcoded from the UML class diagram.

XML DTD

```
<?xml version="1.0" encoding="UTF-8"?>

<!ELEMENT NamedEntity (MorphosyntacticFeatures*, Subelement*)>
<!ATTLIST NamedEntity
        id              ID              #IMPLIED
        type            CDATA           #REQUIRED
        subType         CDATA           #IMPLIED
        formKind        CDATA           #IMPLIED
        source          CDATA           #IMPLIED
        wordForms       IDREFS          #REQUIRED>

    <!--May be more complex, see wordforms and tokens in MAF-->
<!ELEMENT WordForm EMPTY>
<!ATTLIST WordForm
        id              ID              #REQUIRED
        string          CDATA           #REQUIRED>

   <!--To record information like grammaticalGenre, for intance. -->
<!ELEMENT MorphosyntacticFeatures EMPTY>
<!ATTLIST MorphosyntacticFeatures
        att             CDATA           #REQUIRED
        val             CDATA           #REQUIRED>

   <!--To record sub-field elements like given name vs family name -->
<!ELEMENT Subelement EMPTY>
<!ATTLIST Subelement
        type            CDATA           #REQUIRED
        wordForms       IDREFS          #REQUIRED>
```

References

[1] Ehrmann M. 2008 Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation, PhD thesis University Paris 7.

[2] Rumbaugh J., Jacobson I., Booch G. 2004 The unified modeling language reference manual, second edition, Addison Wesley.

[3] The ISO-TC37 data category registry is located at www.isocat.org on October 2009.

[4] The International Press Telecommunications Council (IPTC) newscodes registry is located at www.iptc.org/cms/site/index.html?channel=CH0102 on October 2009.

[5] The Message Understanding Conference (MUC) web site is located at www-nlpir.nist.gov/related_projects/muc on October 2009.

[6] The Automatic Content Extraction (ACE) web site is located at www.itl.nist.gov/iaui/894.01/tests/ace on October 2009.

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

[8] Sekine S., Nobata C. 2004 Definitions, dictionaries and tagger for Extended Named Entity Hierarchy, LREC-2004, Lisbon.

[9] Francopoulo G., Demay F. 2011 A Deep Ontology for Named Entities. International Conference on Computational Semantics, Interoperable Semantic Annotation workshop. Oxford.

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

[11] Poibeau T. 2003 The Multilingual Named Entity Recognition Framework, EACL-2003, Budapest.