

FROM LINGUISTICS TO ONTOLOGIES

The Role of Named Entities in the Conceptualisation Process

Nouha Omrane, Adeline Nazarenko and Sylvie Szulman

Laboratoire d'Informatique de Paris Nord, Université Paris 13 & CNRS (UMR 7030), 99 av. J.-B. Clément, 93430
Villettaeuse, France
firstname.name@lipn.univ-paris13.fr

Keywords: Ontology design, named entity, conceptualisation.

Abstract: Ontologies that have been built from texts can be associated with lexical information that is crucial for the semantic annotation of texts and all semantic search tasks. However, the entire process of building ontologies from texts cannot be fully automated and it is important to guide the knowledge engineer during the building process. This paper presents an enriched version of TERMINAE, which is a text-based methodology for ontology design. It combines a fact-based approach of modeling with the more traditional concept-centric one. We show that named entities can be used to enrich an existing ontology and to bootstrap the acquisition process. In other words, named entities are used for the conceptualisation of ontologies and not only for their population. This approach is illustrated on two use-cases based on policy documents and evaluated by measuring the *Precision* and *Recall* of the resulting ontologies with respect to pre-existing ontologies independently built by domain experts.

1 INTRODUCTION

As defined by (Studer et al., 1998), an ontology is a formal, explicit specification of a shared conceptualisation. Specialized texts are rich sources of information and they are more widely available and exploitable than domain experts who often do not have much time for interviews and are hardly conscious of their own knowledge. Another advantage of building ontologies from texts concerns the future exploitation of the ontology. If the ontology is to be used to annotate documents, for information retrieval or document indexing, it is important to link documents to ontologies and this can be done by recycling the information captured in text-based acquisition.

This paper shows how two types of domain specific textual units – terms and named entities, extracted from texts using natural language processing (NLP) tools – can help the design of ontologies and more specifically their conceptualisation. Beside methods relying on the terminological analysis of texts for the building of ontological models (Terminological Box, or T-Box, in Description Logic), many works have tried to exploit named entity recognition for the population of these ontologies (Assertional Box, or A-Box) (Maynard et al., 2008). However, we argue that named entities are also worth consid-

ering for conceptualisation and that taking these textual entities into account yields to improve T-Boxes (hereafter ontologies).

The TERMINAE methodology (Aussenac-Gilles et al., 2008) is enhanced by taking named entities into account in addition to terms¹. In this paper we focus on the creation of concepts and instances. Many of our conclusions can be extended to the creation of properties and instance properties, but this point is not developed here.

The resulting methodology is illustrated in the context of business rules and policy modeling, where knowledge engineers have to design ontologies, which are then used as conceptual vocabularies for the annotation of source documents and the design of business rules.

Section 2 presents the current approaches in text-based ontology building. Although named entities and terms are traditionally given distinct roles, Section 3 explains that terms and named entities can be exploited in a unified way and shows how the TERMINAE methodology has been enriched with the output of named entity recognition tools. Section 4 illustrates the approach on two use-cases. The last section

¹The corresponding TERMINAE tool has been improved accordingly (<http://www-lipn.univ-paris13.fr/~szulman/logi>) but we focus on the methodological aspects here.

evaluates our approach.

2 TEXT-BASED APPROACHES TO ONTOLOGY BUILDING

Text-based ontology building methods differ on the degree of automaticity they claim to achieve and in the kind of textual elements they rely on (words, terms and named entities²).

Towards semi-automatic methods Distributional approaches rely on the hypothesis that clustering words on the basis of their contextual distribution yields to semantic classes that can then be interpreted as concept. This approach has proved to be more reliable on specialized acquisition corpora and is often referred as "ontology learning", but it can hardly be considered as automatic. In the OntoLp plug-in of the Protege ontology editor, human intervention is required for filtering out words and words semantic clusters (Lopes and Vieira, 2009). Text2Onto (Cimiano and Völker, 2005) exploits various NLP tools to build a draft ontology but the resulting ontology needs to be manually edited by the knowledge engineer afterwards (Wang et al., 2006). Systems like ASIUM (Faure and Nédellec, 1999) or Doodle-OWL (Morita et al., 2008) directly implement ontology design as an incremental or interactive process. Even if word similarities seem to be helpful in the modeling process, word classes are seldom directly exploitable as ontological concepts.

Terminological Approach Among human-centered approaches, the importance of domain terminology for the building of domain specific knowledge is acknowledged for years (Meyer et al., 1992; Aussenac-Gilles et al., 1995) and TERMINAE methodology relies on a terminological analysis. Terminology is often defined as the body of words or terms relating to a particular subject, field of activity or branch of knowledge³. If we consider that each domain of knowledge has its specific sublanguage(s), the terms form the vocabulary associated to such a sublanguage. A term is often a multi-lexical unit rather than a simple word since compounds usually have a more precise, more specialized and less ambiguous meaning than words. Another important property of terms is their relative stability. Even if a term may have variant forms, this variability is lower

²Much work has also been done for relation extraction, but, as mentioned above, we do not focus on properties here.

³Cf. *Collins Dictionary*.

than for plain words which can often substitute for other words or phrases.

These domain specific and stable textual units are useful for ontology building but termhood cannot be defined on purely linguistic or statistical ground. Linguistic and surface clues help to restrict the set of candidate terms but the final selection requires human expertise and a general understanding of the domain.

The Role of Named Entities Named entities are another type of domain specific textual units. They have a referential meaning and refer to well identified domain entities.

Although various types of mentions refer to defined entities (proper names as "American Airlines", pronouns like "it" and definite descriptions such as "this company") (LDC, 2004), we focus on proper names that are easier to identify in documents. Named entity recognition (NER) tools have been designed to locate proper names in texts (persons, locations and organisations but also temporal expressions or measure units) and to associate a semantic type to them (e.g. PERSON, LOCATION) (Nadeau and Sekine, 2007).

Named entities are traditionally exploited in ontology engineering but for populating the instance level of ontologies rather than for their conceptual structuring. (Magnini et al., 2006; Giuliano and Gliozzo, 2008) try to (semi-)automatically enrich the knowledge base (A-Box) associated to a given ontology by discovering new concept instances. Recognising a named entity of type T yields to the creation of an instance of the concept T .

3 A COMBINED METHOD FOR BUILDING ONTOLOGIES FROM TEXTS

Assuming that texts do not contain all useful information and that ontology design depends on the application for which the ontology is to be used, TERMINAE proposes an interactive approach where the knowledge engineer relies on domain specific textual units to model an ontology (Section 3.1). Whereas terms are traditionally used in conceptual design and named entities for ontology population, we argue that both types of textual units can be exploited for the conceptualisation (Section 3.2) and its bootstrapping (Section 3.3). This approach is embodied in a new version of TERMINAE tool (developed as an eclipse application and available as a NeonToolkit plugin).

3.1 TERMINAE Layered Approach

The TERMINAE method decomposes the building process into three main steps – the extraction, normalisation and formalisation (or ontological) steps. The overall process is represented on Figure 1⁴. At each level, the knowledge engineer has to select the relevant items and organise them. This process is helped by the previous terminological analysis of the text which is automatic (step 1), and by the guidelines of the method embodied in the interfaces of TERMINAE (steps 2 and 3).

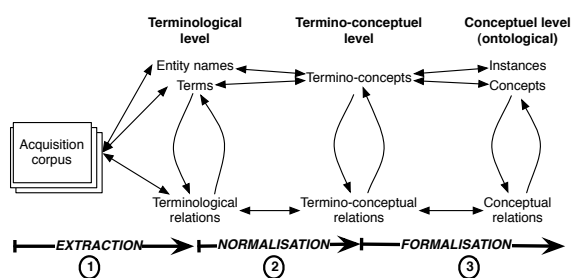


Figure 1: Abstracting a conceptual model out of text: the TERMINAE layered approach.

At the terminological level, the knowledge engineer identifies the textual units of the acquisition corpus that seem to be relevant for the domain and use-case to model. This step relies on NLP tools known as term extractors and named entity recognition tools⁵, but the knowledge engineer has to rework manually the results to rule out ill-formed units, filter out the irrelevant ones and cluster variants or synonyms under a single canonical representative. Only a rough analysis is usually done at that level but the selection and clustering go on during the next building step.

At the termino-conceptual level, the list of relevant textual units is normalised into a network of termino-concepts structured by the BT/NT (broader than/ narrower than) links. Each group of synonymous terms must be clustered and associated to a unique termino-concept. Ambiguous terms must be disambiguated in such a way that each relevant meaning is represented as a distinct termino-concept.

The third building step formalises the network of termino-concepts into an ontology. Several important questions regarding the concept/instance/property distinction and the inheritance hierarchy must be tackled at the conceptual level.

⁴We focus hereon the upper level of the figure.

⁵In the reported experiments, YaTeA term extractor (<http://search.cpan.org/~thamon/Lingua-YaTeA-0.5>) and ANNIE NER tool (<http://gate.ac.uk/sale/tao/splitch6.html#chap:annie>) have been used.

In TERMINAE, the correspondence between the levels is established through the creation of terminological and termino-conceptual forms (see Section 3.2). The termino-conceptual network can be exported as a SKOS file⁶. The conceptual level is realized within Neon Toolkit Ontology editor⁷ which is plugged in our tool.

3.2 Named Entities Conceptualisation

The core task of ontology building is the conceptualisation step that consists in choosing, structuring and defining the conceptual elements of the domain model. The introduction of named entities at the terminological level raises the question of their role in the conceptualisation process, which mainly relies on the knowledge engineer's understanding of the domain and aimed application.

Our approach differs from the works on ontology population which tend to automatically derive the ontological types of conceptual elements (instances vs. concepts) out of their linguistic type (named entities vs. concepts). TERMINAE allows to link any type of conceptual elements to any type of textual unit through the termino-concepts (terms and named entities to termino-concepts and termino-concepts to concepts, instances or conceptual relations), as shown on Figure 1. The distinction between the T-Box and the A-Box levels does not exist at the termino-conceptual level. The relevant linguistic units have to be linked to termino-concepts or termino-conceptual relations at the termino-conceptual level. The conceptualisation choices come after, in the transition from the termino-conceptual level to the ontological one.

The first conceptualisation step is handled in TERMINAE through the creation of terminological forms for any textual unit that is considered as relevant for the domain to model. A terminological form describes the properties of a term or named entity and associates it to one or several termino-concepts, one for each relevant meaning. Figure 2 presents the form of *Sapphire*, the textual unit that had been selected in the list on the left side⁸. The lexical information indicates whether it has been extracted as a term, a named entity or both. By browsing the list of its occurrences, the knowledge engineer decides whether this unit is an important one or not (selection) and how many of its meanings are relevant for the use-case (ambiguity analysis). For each relevant meaning, a termino-concept is created, the knowledge en-

⁶<http://www.w3.org/2004/02/skos/>

⁷http://neon-toolkit.org/wiki/Main_Page

⁸*Sapphire* is a specific category of travellers in the Air-line use-case (see Section 4).

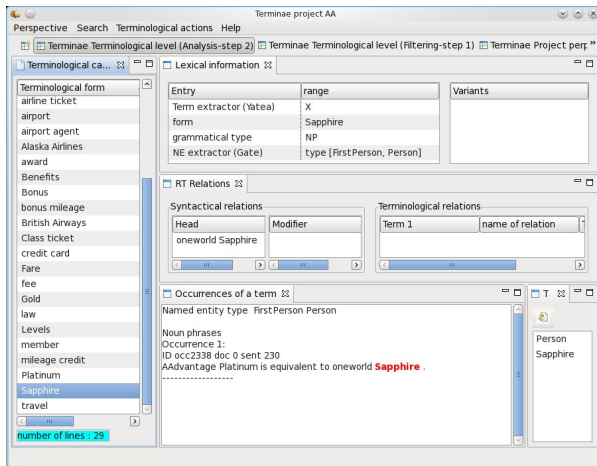


Figure 2: Terminological form of the named entity *Sapphire*

gineer can give a natural language definition, associate synonyms (clustering) and select the most representative occurrences for the newly created termino-concept. Figure 3 presents the termino-concept *Sapphire*. *Sapphire* and *AAdvantage platinum* are considered as synonyms.

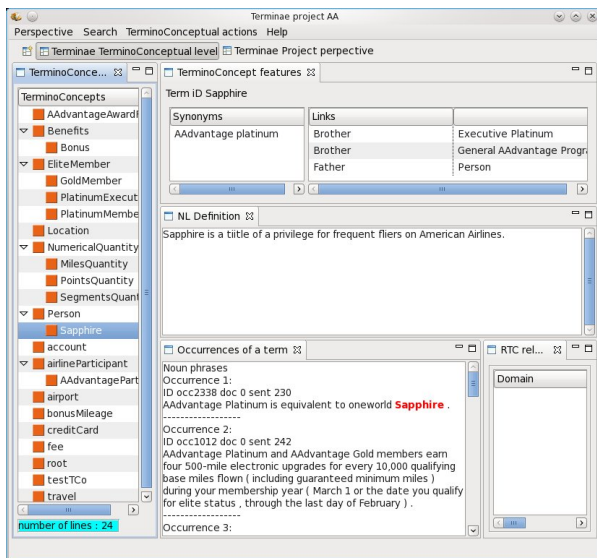


Figure 3: Termino-concept *Sapphire*

The case of named entities is specific as they are associated to a semantic type by NER tools (PERSON in the case of *Sapphire*). This semantic type may also be relevant for the domain to model. As they have no linguistic counterpart, semantic types are represented directly at the termino-conceptual level with no occurrence attached to them even if they remain associated to their named entities as the fathers of the corresponding termino-concepts (see the hierarchical list of termino-concepts on the left side of Figure 3). The

terminological form of the named entity *Sapphire* is thus associated with the termino-concepts *Sapphire* and *Person*, the latter being built from the semantic type PERSON given by the NER tool.

3.3 Bootstrapping Conceptualisation

Mining long lists of terms, named entities, termino-concepts or concepts is difficult. The broad-first search is not practicable as the terminological analysis cannot be completed without a view of the intended ontology. Conceptualisation rather relies in a deep-first search strategy but the problem is to choose the textual units and concepts to start with for bootstrapping the conceptualisation process.

We propose here a fact-based approach. Starting with the named entities that are expected to refer to core domain specific entities leads to create the concepts the named entity belong to, the relations they hold and more generally to model what is said about them (in the same sentences or textual areas).

4 EXPERIMENTS

We consider two use-cases dealing with regulations (loyalty program, EU directives). The resulting ontologies are to be used for the modeling and formalization of the rules that are expressed in written policies. The final rule bases and underlying ontologies are intended to be used in Business Rule Management Systems in the FP7 ONTORULE project. These experiments show that the named entities are important to take into account even in texts that do not contain a lot of them.

The ontology building scenario differs in the two experiments reported below. In the first one, the named entities are exploited to enrich an ontology that we had previously built on the basis of terms only. The second experiment shows how useful named entities are for bootstrapping the conceptualisation.

4.1 Airline use-case

In the first experiment, the ontology is built out of a document explaining the mileage policy of American Airlines (AA) to customers (5 744 words).

Having built an initial ontology from a list of 973 terms, we considered the 67 named entities identified in the corpus and some of them appeared as highly relevant for the domain.

For instance, *Central America* refers to a specific type of airports that plays a specific role in the attribution of miles. The underlying notion had

not showed up in our initial term analysis but the named entity list has brought it to light and we finally modelled it as a concept in the ontology (with various airports as instances). The named entities referring to airline companies (*e.g.* *Japan airways*) were modelled as instances of the concept **AAdvantage Airline Participant** that we had already created from the term *AAdvantage participant* (airline companies participating in the “AAdvantage” program). The semantic type ORGANIZATION lead us to create a concept **Organization** as a father concept of **AAdvantage Airline Participant**. The named entity analysis also revealed the importance of *Sapphire* and *Ruby* which had initially been considered as noisy terms but actually refer to customer categories.

Detecting these named entities allowed us to better understand the airline policy underlying domain and lead us to revise the initial ontology. The final one contains 137 domain concepts.

4.2 Regulation use-case

The Regulation use-case is based on EU directives describing policies on seat belt control procedures (3 704 words).

We started modeling the domain by considering the named entities having PERSON, DATE, PERCENT and UNKNOWN as semantic types⁹ and by exploring their contexts¹⁰ as shown on Figure 4.

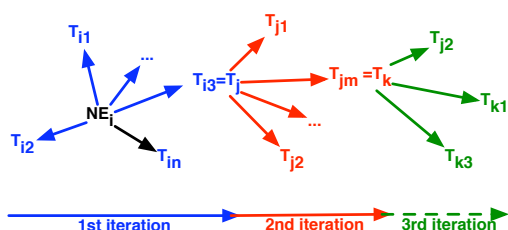


Figure 4: Exploration strategy. Starting with a given named entity (NE_i), the knowledge engineer analyses the surrounding terms (T_{i1} to T_{in}) during a 1st iteration before further exploring the list of terms during additional iterations.

Bootstrapping the conceptualisation starting with the selected named entities (1st iteration) helped identifying relevant domain terms during the first regulation browsing. For instance, the names of the tests (*e.g.* *Calibration Test*) used to check the conformance of seat belts are emphasised. Named entities of PERCENT type help the knowledge engineer to detect im-

⁹A quick analysis showed that the named entities of other semantic types (ORGANISATION and LOCATION) were not relevant for the domain to model and the aimed application.

¹⁰The occurrence and the two surrounding sentences.

portant properties of domain concepts (mainly testing conditions). The UNKNOWN named entities often correspond to specific elements of the vocabulary of safety belt procedures such as categories of vehicles ($M1$, $N1$) or specific belt positions (*Point A*, *Point C*).

In all, we created 53 domain concepts during the bootstrapping step.

5 EVALUATION

We compare the ontologies built for each use-case with those that were previously built by an expert who did not start from texts (hereafter “expert ontologies”). The comparison is based on the labels of concepts or termino-concepts. We use *Precision* and *Recall*¹¹ measures to evaluate the resulting ontologies with respect to the pre-existing ones.

For each use-case, we created a draft ontology, through a single iteration of the whole building process. The goal is to evaluate the role of named entities either for enriching a term-based ontology (Airline use-case) or in the bootstrapping strategy (Regulation use-case).

In the Airline use-case, we compare a first ontology built without taking named entities into account and a revised version enriched with named entities with respect to the expert one. There is no major difference between the first and enriched versions of the ontology (from 0.828 to 0.830 for *Precision* and from 0.675 to 0.720 for *Recall*), due to the small number of named entities. However, taking the named entities into account has led to re-structure the first ontology (11.5% of the existing concepts have been redefined), enrich it (the revised version is 40% larger) and populate it (45 instances have been added). In all, 60% of the named entities have been introduced in the ontology in a way or another.

The Regulation use-case presents a high *Precision* (0.742) but a low *Recall* (0.393): the ontology contains mainly relevant domain concepts but only partially covers the domain described by the expert. The concepts that are missing in our ontology belong either to the core level of the ontology or to specific sub-types of existing concepts (*e.g.* categories of vehicles). The former ones are concepts that are not directly expressed in texts. Detecting the latter would require a finer-grained analysis and additional iterations of the conceptualisation process, but most of the related terms occur in the contexts of the mentions of existing concepts.

¹¹ $P = \frac{\text{number of relevant termino-concepts } RTC}{\text{number of termino-concepts } TC}$
 $R = \frac{\text{number of relevant termino-concepts } RTC}{\text{number of concepts in the expert ontology } C}$

Both use-cases show that named entities help the detection of relevant domain concepts. Since the list of named entities extracted is much smaller than the list of terms (more than 10 times smaller in our use-cases), it is interesting to rely on named entities as a starting step, even if the list of textual units must then be further explored to enrich the draft ontologies based on named entities.

6 CONCLUSION

This paper shows how text-based ontology building methods can be enriched by taking specific textual units into account – named entities as well as terms – and explains how named entities can be used in the conceptualisation task. We show that they can be used either to enrich an ontology (building new concepts, their properties, re-structuring and populating existing concepts) or for bootstrapping the conceptualisation step and identifying relevant domain terms.

This combined approach, which is implemented in the new TERMINAE tool, is illustrated on two use-cases. Even if named entities are not as numerous in policies as in press articles for instance, they are important to take into account in the conceptualisation process because they point out critical domain elements that are important to integrate in a conceptual model.

ACKNOWLEDGMENTS

This work was realised as part of the FP7 231875 ONTORULE project (<http://ontorule-project.eu>). We are grateful to American Airline who is the owner of one of our working corpora.

REFERENCES

Aussenac-Gilles, N., Bourigault, D., Condamines, A., and Gros, C. (1995). How can knowledge acquisition benefit from terminology? In *Proceedings of the 9th Knowledge Acquisition Workshop*.

Aussenac-Gilles, N., Despr'es, S., and Szulman, S. (2008). The terminae method and platform for ontology engineering from texts. In Buitelaar, P. and Cimiano, P., editors, *Bridging the Gap between Text and Knowledge*, pages 199–223. IOS Press.

Bannour, S., Audibert, L., and Nazarenko, A. (2011). Mesures de similarité distributionnelle entre termes. In *22èmes journées francophones d'Ingénierie des connaissances*.

Cimiano, P. and Völker, J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In *Proc. of the 10th Int. Conf. on Applications of Natural Language to Information Systems*, pages 227–238.

Faure, D. and Nédellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning: the system asium. In et R. Stude, D. F., editor, *Proc. of the 11th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW'99)*, pages 329–334. Springer-Verlag.

Giuliano, C. and Gliozzo, A. (2008). Instance-based ontology population exploiting named-entity substitution. In *Proc. of the 22nd Int. Conf. on Computational Linguistics (Coling 2008)*, pages 265–272, Manchester.

LDC (2004). Ace (automatic content extraction) english annotation guidelines for entities. Livrable version 5.6.1 2005.05.23, Linguistic Data Consortium.

Lopes, L. and Vieira, R. (2009). Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *Electronic Journal of Communication, Information and Innovation in Health*, 3(1):72–84.

Magnini, B., Pianta, E., Popescu, O., and Speranza, M. (2006). Ontology population from textual mentions: Task definition and benchmark. In *Proc. of the OLP2 workshop on Ontology Population and Learning*.

Maynard, D., Li, Y., and Peters, W. (2008). NLP techniques for term extraction and ontology population. In Buitelaar, P. and Cimiano, P., editors, *Bridging the Gap between Text and Knowledge*, pages 199–223. IOS Press.

Meyer, I., Skuce, D., Bowker, L., and Eck, K. (1992). Towards a new generation of terminological resources : an experiment in building a terminological knowledge base. In *Proc. of the 15th Int. Conf. on Computational Linguistics (COLING'92)*, pages 956–960, Nantes, France.

Morita, T., Fukuta, N., Izumi, N., and Yamaguchi, T. (2008). Doodle-owl: Interactive domain ontology development with open source software in java. *IEICE Transactions on Information and Systems*, (4):945–958.

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

Studer, R., Benjamins, V. R., and Fensel, D. (1998). *Knowledge Engineering: Principles and Methods*, volume 25, pages 161–197.

Wang, Y., Volker, J., and Haase, P. (2006). Towards semi-automatic ontology building supported by large-scale knowledge acquisition. In *AAAI Fall Symposium On Semantic Web for Collaborative Knowledge Acquisition*, volume FS-06-06, pages 70–77, Arlington, VA, USA. AAAI, AAAI Press.