Combining terms and named entities for modeling domain ontologies from texts

Nouha Omrane
University of Paris 13
99 av JB Clement
93430 Villetaneuse, France
nouha.omrane@lipn.univ-paris13.fr

Adeline Nazarenko
University of Paris 13
99 av JB Clement
93430 Villetaneuse, France
adeline.nazarenko@lipn.univ-paris13.fr

Sylvie Szulman
University of Paris 13
99 av JB Clement
93430 Villetaneuse, France
sylvie.szulman@lipn.univ-paris13.fr

ABSTRACT

Building ontologies from plain texts is still a research issue. This process cannot be fully automated but natural language processing and methodological guidelines can help the knowledge engineer's task. In this paper we present TERMINAE and show through the analysis of three different experiments on policy documents how the initial terminological approach can be guided by taking named entities into account.

Keywords. Ontology acquisition from texts, terms, named entities, conceptualization.

1. INTRODUCTION

Specialized texts are rich sources of information and they are more widely available than domain experts who often do not have much time for interviews and are hardly conscious of their own knowledge. There exist two main text-based approaches for designing ontologies.

The first "ontology learning" approach [3] relies on distributional analysis of large acquisition corpora. It is considered as an automatic one, even if the resulting ontology needs to be manually edited afterwards. The second approach is based on the terminological analysis [1] of the text. It is less automated than the previous one but is useful for applications where ontologies need to be carefully designed.
This work is part of a project aiming at modeling business rules expressed in written policies. In this context, where domain ontologies are used as conceptual vocabularies for the writing of the rules of various use cases, the terminological approach is prefered given the typical size of policies (medium size specialised corpora) [1] and the expected quality of the ontologies. In the terminological approach, terms of a domain form the domain specific vocabulary and, as such, serve as a bootstrap for ontology design. Named entities are another type of domain specific textual units that refer to well identified domain entities. They are traditionally exploited in ontology engineering but for populating the instance level of existing ontologies. The originality of the proposed method comes from the fact that it exploits both types of textual units to bootstrap the conceptualization process itself. Our approach is a terminological fact-based one that is embodied in a revised version of TERMINAE tool [2], which now takes named entities into account in addition to terms. Section 2 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 rules. The last section presents three different experiments exploiting named entities in the ontology building process.

[0]This work was realized as part of the FP7 231875 ONTORULE project. We thank American Airline and ArcelorMittal who are the owners of our working corpora.
[1]Typically, from 5 to 500 thousands of words

2. A COMBINED METHOD FOR BUILDING ONTOLOGIES FROM TEXTS

The TERMINAE text-based acquisition method decomposes the acquisition process into three main levels – the terminological, termino-ontological and conceptual (or ontological) levels – which are built on top of each other, the corpus playing the role of ground level. The transition from text to ontology must actually be mediated. Ontologies cannot be "extracted" as such from texts, because conceptual models (or ontologies) and texts are different in nature. At each level, the knowledge engineer has to select the relevant items and to organise them. This process is helped by the previous terminological analysis of the text, which is automatic, and guided by the method embodied in the interfaces of TERMINAE tool. The overall process is represented on Figure 1. In this paper, we focus on the upper part of it. At the linguis-

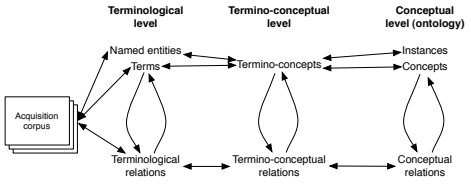

Figure 1: Abstracting a conceptual model out of text: the layered approach.

tic level, the user has to extract from the acquisition corpus

the textual units that seem to be relevant for the domain and use case to model. This step relies on NLP tools known as "term extractors", as well as "named entity recognizers" that extract named entities and their semantic types. The user has to revise the extracted elements and to turn the list of relevant units into a list of termino-concepts. In that process, the linguistic output is normalised, which is a way to abstract the future domain model from the textual wording and linguistics. The third acquisition step of TERMINAE methodology consists in formalising the list or network of termino-concepts into an ontology. The core task of ontology acquisition is the conceptualization step that consists in choosing, structuring and defining the conceptual elements of the domain model. In this step, named entities are generally neglected. On the contrary, we consider these textual units and their semantic types from the beginning of the conceptualization phase in the same way as we do for the terms. It is not because they are identified as named entities by NLP tools that they must necessary be turned into instances. In some cases the named entities might model concepts. The underlying modeling choices depend on the corpus and use case that are considered.

The next section illustrates the various bootstrapping approaches in the context of policy modeling taking into account the specificities of policy documents in which passages expressing rules deserve specific attention.

3. EXPERIMENTATIONS

We consider three use-cases, each one dealing with a specific type of regulations (loyalty program, decision process, rules of a game). The resulting ontologies are to be used for the modeling and formalization of the rules that are expressed in written policies. The acquisition scenario is not the same in the three 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 case aimed at adding linguistic information to an existing ontology and at enriching it with information coming from the acquisition corpus. In the third experiment, the named entities are realy used to bootstrap the conceptualization. Even if the policy corpora do not contain numerous named entities, the three experiments show that the named entities are important to take into account.

In the first experiment, the ontology is built out of a document of American Airlines (5, 300 words), which explains mileage policy to customers. In this use case, taking the named entities into account yields to enrich and partially populate the ontology. Compared to the initial ontology of 130 concepts, 7 new concepts and 45 instances have been added. 15 of the existing concepts have also been redefined. Except for cities which were not interesting for the use case, all the named entities (76) have been introduced in the ontology in some way. The second use-case deals with the galvanization process and the rules dealing with the assignement of a product (coil)(3, 562 words) : depending on various quality criteria, a coil can be assigned to the order (delivered to the customer), repaired or thrown away. We started modeling the domain from an existing core ontology of 12 concepts. The goal was to associate textual units to existing concepts[2] and to enrich the structure of the ontology with entities which have been found in the text. We exploited the 663 terms and 105 named entities respectively extracted by YaTeA[3] and Gate[4]. Taking named entities into account helped to understand the details of the assignement process and to identify the relevant conceptual properties. In the third experiment, we had no preexisting information and we exploited the named entities to bootstrap the conceptualization process, in a fact-oriented approach. We started with a French "Rules of Golf" corpus[5] (112,898 words) which describes the rules and conditions according to which a golf player must replay, loose points or quit the game. YaTeA and Gate respectively extracted 3, 711 terms and 350 named entities. In this use case where the term list were too long to be studied in detail, the analysis started with named entities which underlined some core domain elements and was progressively extended to the related terms and their interrelations. These three experiments aimed at building ontologies out of written policies. Named entity recognizers bring into light textual units that are not identified as terms but which nevertheless refer to crucial domain elements and guide the conceptualization work. Even if the "populating" hypothesis does not hold – named entities can be modeled as concepts as well as instances –, named entities favour a fact-oriented approach, which counterbalance purely terminological analyses.

[2]for the further semantic annotation of additional documents

4. CONCLUSION

This paper shows how text-based ontology acquisition methods can be enriched by taking all types of domain specific textual units into account, named entities as well as terms, and explains how named entities can be used in the conceptualization task.

This combined approach, which is implemented in the TERMINAE tool, is illustrated on three different experiments that all aim at building ontologies for the modeling of rules. The written policies do not have as many named entities as press articles for instances, but we have shown that they support a fact-based modeling approach that is complementary to the terminological one, which is more concept-oriented. Even when they are represented as instances at the conceptual level, named entities point out critical domain specific elements that are important to integrate in the conceptual structure in a form or another.

5. REFERENCES

[1] Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer (2006)

[2] Aussenac-Gilles, N., Despres, S., Szulman, S.: The TERMINAE Method and Platform for Ontology Engineering from texts. In Buitelaar, P., Cimiano, P., eds.: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text. IOS Press (janvier 2008) 199–223

[3] Lopes, L., Vieira, R.: 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) (2009) 72–84

[3]http://search.cpan.org/%7Ethhamon/Lingua-YaTeA-0.5/

[4]http://gate.ac.uk/

[5]This public document is available on http://www.ffgolf.org/