Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security)

Olena Orobinska[1], Jean-Hugues Chauchat[2], Natalya Sharonova[3]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

olena.orobinska@univ-lyon2.fr[1],
jean-hugues.chauchat@univ-lyon2.fr[2], sharonova@kpi.kharkov.ua[3]

Abstract. We propose a hybrid, semi-automatic approach that uses the intersection of semantic classes of nouns and verbs built on the domain lexicon and builds kernel ontology from a list of initial concepts and then completes this kernel ontology by new entities detected in a large corpus of texts of international standards of Radiological Safety. The results confirm the important role of initial linguistic modeling and show that the external lexical resources available online can contribute effectively to the resolution of the problem of lexical disambiguation.

Keywords: ontology learning, text processing, semantic analysis, terms extraction

1 Introduction

In the community of ontology researchers and computer scientists, is recognized that the construction of an ontology involves some steps. It's common to start with the installation of a kernel ontology which includes either the simple enumeration of the denominations of the concepts or, in addition, a hierarchy of these concepts. The kernel ontology is used to extract new candidates.

We propose to include the extraction of semantic relations to the first step. In other words, we propose to give the same importance to the concepts and the relations that joins them. Thus, we start with the anticipated conceptualization of the modeling domain and, at the same time, the anticipated linguistic modeling of the input corpus and introduce the notion of predicative framework.

2 Semantic and Linguistic Modeling of Kernel Ontology

The specific terminology of a certain domain is univocal. Thus the denominations of the concerned phenomena, physical quantities, units of measurement are strictly

defined and listed in the specialized glossaries. It does not vary much in technical and scientific texts. On the other hand, a concept always presents a class of objects possessing similar properties. There are two ways of defining a concept: either by its intention, i.e. the explicit definition restricting its properties, or by its extension, i.e. by the enumeration of objects that posses its characteristic properties. We have chosen the representation of concepts by their extensions.

Hence, the following definition of the kernel ontology.

Definition. The kernel ontology is the combination of the list of semantic classes of names; each class corresponds to the extension of a concept, and the predicative framework modeling their semantic relations.

For the construction of an ontology we propose the following operating mode:

1. In consultation with the experts, de ne a limited list of general terms and categories of semantic relations between these terms.
2. Taking each term as a reference for a concept, grouping around them its synonyms to constitute the semantic classes representing the concepts through their semantic extensions.
3. Form the predicative framework in the form of the set of lexical-semantic classes of verbs.
4. Apply the predicative framework for the extraction of new candidates-terms to populate the ontology.

Note that the order of items 2 and 3 is exchangeable.

2.1 Initial List of Concepts

The selection of the initial concepts was carried out in 4 steps.

1. At the beginning, we extracted from the two corpuses, French and Russian, the 100 ~~–b~~est" candidates-terms according to the TF-IDF index.
2. To select the general concepts of the domain, we used the RISK framework, which summarizes the situations related to risks of any kind.
3. The final validation by the expert allowed to retain a list of ten words, these becoming the initial denominations of the concepts. This list includes the following terms (in French and Russian): *damage*, *exposure*, *control*, *personnel*, *population*, *protection*, *radiation*, *risk*, *safety* and *source*.

We perform a first grammatical analysis to recover in the corpus the pairs of the type (*w*, *v*), where *w* is the name and *v* is an ~~–c~~haracteristic" verb in the same sentence. In the complete list of all noun-verb pairs, we keep those that contain predefined terms or their synonyms suggested by the dictionary. A module in Java has been written for this step.

The evaluation of the synonyms of the initial terms was carried out according to the FCA method: two names are considered to be true synonyms if they are associated with the same characteristic verbs. In order to select the characteristic verbs which form the formal context of each concept, we proposed to measure the degree of association between each general term and each of the verbs associated with it in the corpus with the coefficient **K** (1) that is the product of the Mutual Information (MI) and the Jaccard Coefficient.

$$K = MI(c_i, v_j) \cdot JaccardCoefficient(c_i, v_j) \tag{9}$$

where

$$MI(c_i, v_j) = |W| \cdot \log TF(c_i, v_j)/(TF(c_i) \cdot TF(v_j))$$

and

$$JaccardCoefficient(c_i, v_j) = TF(c_i, v_j)/(TF(c_i, \bar{v}_j) + TF(\bar{c}_i, v_j))$$

2.2 Predicative Framework

Relationships contribute to the construction of an ontology in the same way as concepts.

Definition. By predicative framework we mean the set of lexical indices which explain the relations between the concepts and make it possible to detect them in the corpus.

We focus on verbs as they are the main predicative agents: each semantic relation category corresponds to a certain predicate and each predicate can be realized using several verbs which in this case form a semantic class.

The diversity of the grammatical and lexical means of a language to express the relations between the objects of the real world complicates their emphasis in the texts. One of the most explicit ways of doing this is using verbs. In this method, we use a superficial analysis of sentences to extract the subject-verb-object (SVO) triplets, subject and object being represented by terms designating the concepts. As a rule, the subject is expressed by a nominal group to the left of the verb, while the object is a nominal group to the right of the verb. In the case of a passive construction, these places are reversed. The use of lemmas makes it possible to reduce the sensitivity of the method to this inversion.

In the first step, as for the names in the previous method, we retrieve in the corpus the potential synonyms of the verbs, selected using the CRISCO Dictionary of Synonyms. But this operation is not sufficient to constitute the semantic classes because most verbs are polysemic and because the dictionary does not explicitly distinguish the different types of semantic similarity, notably the hierarchy (or subsumption) and equivalence, which are realized by different predicates and have different properties in logical theory.

The justification for choosing a good criterion to evaluate the semantic similarity of two words is non-trivial [1]. In order to quantify and measure the degree of synonymy between verbs, we tested the Cosinus measure (2).

$$simCos = \frac{V_i^C \cap V_j^C}{\sqrt{|V_i| \times |V_j|}} \tag{10}$$

Here $V_i^C \cap V_j^C$ is the number of co-occurrences of verb v_i and v_j with the same concept; and $|V_i|$ and $|V_j|$ are the co-occurrences of these verbs with the other names

of the corpus.

2.3 Terminology Pattern Method

The working hypothesis of this method is that the domain lexicon can be detected in the specialized corpus using linguistic analysis. By having a list of generic terms and by empirically discovering the frequent syntactic structures in which these terms appear, we can extend the kernel ontology by new terms, forming the taxonomy, [2]. For example, the term dose is part of the lexicon of the Radiation Security domain. Varied terms, such as *effective dose*, *effective collective dose*, etc. are formed around it.

According to [3], terms are formed by hierarchical syntactic structures. And to enrich the kernel ontology, it's possible to use terminological patterns, which we de ne as the morpho-syntactic structure with one of the generic terms at the head of each. Our goal is to establish these patterns. Terminological pat-terns are formed in two ways: from the analysis of the frequencies of syntactic structures in the corpus; then from the syntactic analysis of the terms of the domain glossary. The fragments of sentences that correspond to the patterns are extracted automatically from the corpus and then validated by the expert. By construction, all extracted fragments contain generic terms that form the kernel ontology: one of the generic terms is the radical of each new term. After validation, terms derived from the same root form a partial taxonomy. They are added in ontology as corresponding concepts.

Initially the patterns are N-grams of grammatical tags that have replaced the words in the corpus. We use N-grams varying from 2 to 6 and extract from the corpus all the fragments of sentences corresponding to these N-grams. The selection of potentially relevant patterns was made from the initial list of generic terms.

3 Conclusion

During our work we have proposed and implemented a coherent algorithm for the construction of ontology in the domain of Radiation Security. These include the formation of semantic classes representing concepts and their relationships, the learning of morpho-syntactic patterns and the installation of partial taxonomies of terms.

All methods are integrated, starting from a limited list of general terms, previously defined with the domain expert. The implementation of this approach required the installation of two corpuses specialized in the domain of Radiation Security, in French and Russian, with 1,500,000 and 600,000 lexical units respectively. A broad synthesis on the state of the art preceded the experimental stage. It covers the various aspects of ontology learning: the theoretical foundations of knowledge representation, natural language modeling, the extraction of terms and relations, the conceptualization phase and the panorama of available tools.

The results have been published in 13 national and international journals and proceedings, between 2010 and 2016, including IMS-2012, TIA-2013, TOTH-2014, Bionica Intellecta, Herald of the NTU –KhPI".

References

1. Nokel, M., Loukachevitch, N.: An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri In: Proceedings of 10th International Conference on Terminology and Arti cial Intelligence TIA 2013 pp.69-76. Paris (2013)
2. Orobinska, O., Chauchat J.-H., Sharonova N.: Enrichissement d'une ontologie de domaine par extension des relations taxonomiques a partir de corpus specialis In: Proceedings of the 10th International Conference on Terminology and Arti cial Intelligence TIA 2013, pp.129{137. Paris (2013)
3. Cabre, M.T., Cormier, M.C., Humbley, J.: La terminologie: theorie, methode et applications. Presses de l'Universit d'Ottawa, (1998)