Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction

Wiktoria Golik  
MIG INRA UR1077  
Domaine de Vilvert,  
F-78350 Jouy-en-Josas,  
France  
wiktoria.golik@jouy.inra.fr

Pierre Warnier  
MIG INRA UR1077  
Domaine de Vilvert,  
F-78350 Jouy-en-Josas, France  
pierre.warnier@jouy.inra.fr

Claire Nédellec  
MIG INRA UR1077  
Domaine de Vilvert,  
F-78350 Jouy-en-Josas, France  
claire.nedellec@jouy.inra.fr

Abstract

The automatic population of a termino-ontology is a difficult and challenging task. We propose a text-based ontology extension method that was experimented and evaluated on, for a semantic annotation task in the biomedical domain. It is based on the linguistic analysis of terms and their heads. The head-based method improves both the identification of relevant areas of a termino-ontology and the matching of the corpus terms within these areas.

1 Introduction

Ontology population has seen much advancement in the past several years. It is an important area of research, in particular in the semantic annotation domain. There are several methods that specifically target ontology enrichment from text. All of these methods strive to be as automatic as possible. Some of them focus on the ontology structure in order to automatically infer semantic relationships (hyponyms, meronyms) (Euzenat, 2007). Others are supported by context analysis such as distributional semantics (Grefenstette, 1994) or patterns (Hearst, 1992).

Below we describe a text-based ontology extension method that we experimented on, for a semantic annotation task in the biomedical domain. Given a corpus and a domain specific termino-ontology, its primary aim is to identify the terms of the termino-ontology that are semantically close to the corpus terms. Further, it is used to annotate the corpus terms with the conceptual information provided by the termino-ontology. Conversely, the method will also be used for the automatic identification of relevant areas of the termino-ontology, which are semantically related to the given corpus terms. This mapping will be used for a semi-automatic extension of the selected areas in the termino-ontology with the corpus terms.

For relating corpus terms to termino-ontology terms, the method targets their internal structure. It therefore belongs to the class of linguistic methods based on the morpho-syntactic analysis of corpus terms (Jacquemin & Tzoukermann, 1999). The core of the method is based on the analysis of terms, their heads and the degree of head similarity (Hamon & Nazarenko, 2001). Our method is inspired by MetaMap (Aronson, 2001) which tags biomedical corpora with the UMLS Metathesaurus using syntactic analysis that takes into account lexical heads of terms.

The method has been successfully evaluated on the event extraction task of the BioNLP 2011 Bacteria biotope shared task (Bossy et al., 2011). The prediction of event arguments is done by automatically tagging corpus terms using termino-ontologies. Our linguistics-based approach achieves better results than shallow mapping methods.
2 Method

The BioNLP 2011 Bacteria biotope (BB) shared task consists of identifying bacteria and their locations in scientific documents. The locations belong to eight types to be predicted: Host, Host-part, Geographical, Food, Medical, Soil, Water and Environment. The participant system then has to relate bacteria to locations by a localization relation. Some locations (e.g. Geographical or Host) can be identified using named entity recognition. Contrarily, other types are more difficult to predict, since they can refer to any physical matter and are also subject to deep morpho-syntactic variations. They are noun phrases with adjectival and noun modifiers, verbal and prepositional complements. In order to overcome both the high degree of morphological variation and the incompleteness of the available lexicon, we experiment using a method based on the comparison of terms extracted from the corpus and termino-ontology terms. More precisely, the method identifies the semantically closest terms among them.

Our method applies to termino-ontology resources (TORs), defined as lexicalized ontologies. The terminological level consists of classes of canonical terms and synonyms with their syntactic properties. Each terminological class is attached to an ontology concept. We experimented using two different TORs of similar size, each containing approximately 1600 concepts: the Microorganism Biotope Termino-Ontology (MBTO) and the publically available EnvO habitat termino-ontology (Field et al., 2008). The first, focusing on bacteria biotope and phenotype modeling, has been previously developed at INRA. The second is more generalist, but it also targets habitat modeling. Given that the resources have not been created for the purpose of the experiment, we had to associate the BB task types to the MBTO and EnvO concepts. To do so, we took advantage of the hierarchical structure of the ontologies. We manually associated the high level nodes of the location hierarchies to the eight location types. The types of the lower level concepts were then automatically inferred. Local exceptions were manually handled. Importantly, the type assignment was consistent with the two TOR models: the concepts of the same type belong to contiguous areas.

The method relates corpus terms to TOR terms in two stages. The first stage extracts corpus terms using BioYatea. BioYatea is a version of YaTea (Hamon & Aubin, 2006), that we have extended and adapted to the biology domain. BioYatea provides information about the syntactic structure of terms, including the head and its modifiers. Head identification is a crucial point for our location identification and typing method. We consider that the head of a candidate location term is the most informative part and that it conveys the location type information. In most cases the term head is unambiguous with respect to the type.

At the second stage, the method assigns types to the candidate location terms produced by BioYatea. When both the corpus term and the TOR terms share the same head, the corpus term is assigned the type of the union of matching TOR terms. For instance, the corpus term aquatic sediment shares the head sediment with several TOR terms such as lake sediment or spring sediment. They all belong to the Environment area. Therefore, aquatic sediment acquires the type Environment. There are cases where the corpus term head appears in several type areas. For example, the head spinach from the corpus term decayed spinach belongs to three areas, namely Host (diseased spinach) Environment (decayed spinach) and Food (cooked spinach). A rule-based processing disambiguates among the several types by analyzing the term modifiers. Some heads may also be non informative and cannot be used to discriminate the type. These heads have been automatically detected among ambiguous TOR heads and recorded beforehand. If a candidate corpus term has such a head, we recursively search its subterms for an informative head. For instance, the head environment is non informative. Therefore, for the candidate term cool soil environments, the subterm cool soil is considered. Its head soil denotes the type Soil.

BioYatea extracted 1,873 candidate terms from the test corpus of the BB task. Table 1 details the number of candidate terms that the method associates to TOR terms with respect to the different typing strategies. There are two overall patterns. First, the well-adapted MBTO provides higher results for both strategies, exact match versus head match. Second, for both resources head-matching notably increases the term matching rate. This difference is especially marked for EnvO (91%). The results show that the head-matching strategy alleviates the TOR incompleteness, in particular for less adapted resources such as EnvO.
denoted by terms with monosemous heads than with abstract and polysemous heads.

### References


Thierry Hamon and Sophie Aubin. 2006. Improving term extraction with terminological resources. In T. Salakoski et al. (Editors), Advances in Natural Language Processing 5th International Conference on NLP, (Fin- TAL’06):380-387. Springer.

