Towards a Multi-Level Annotation Format for the Interoperability of Automatic Term Extraction Corpora

Nicola Cirillo¹, Daniela Vellutino¹

¹University of Salerno, 132 Via Giovanni Paolo II, Fisciano (SA), 84084, Italy

Abstract

English. The main corpora used as benchmarks in Automatic Term Extraction are represented in different formats. Unfortunately, none of these formats covers the wide range of linguistic phenomena related to terminology. To address this issue, we propose to encode Automatic Term Extraction corpora in RDF using the OntoLex-Lemon and the NLP Interchange Format ontologies. Furthermore, we developed a small Italian corpus on waste management legislation to provide an example of the proposed formalization.


Keywords

Terminology, Automatic Term Extraction, Linguistic Linked Data, OntoLex-Lemon

1. Introduction

Automatic Term Extraction - ATE is an NLP task that involves recognizing terms in specialized corpora. As with most NLP tasks, ATE research benefits from annotated corpora that are employed as training data and evaluation benchmarks. Nevertheless, existing term annotation schemata are far from capturing the complex organization that characterizes the terminology of specialized languages [1, 2, 3]. Most ATE studies, with a few exceptions [4, 5, 6], overlook the complex organization of terms in specialized languages assuming that the terms contained in a corpus belong to a single domain. At best, they draw a difference between domain terms (that belong to the investigated domain) and out-of-domain terms (that belong to different domains). Unfortunately, this assumption is too simplistic since every specialized corpus contains terms from different subject fields. Moreover, in the interest of reusability, researchers who use terminology corpora in their work must be able to define the subject fields of interest according to their needs.

Furthermore, ATE corpora do not adhere to standard formats used to encode terminological data like TermBase eXchange, an ISO standard [7], and OntoLex-Lemon, a W3C standard. This lack of standardization poses interoperability issues and hinders the evaluation of ATE tools. The adoption of a standard format will provide at least three main benefits:

- It will grant the interoperability of termbases. Therefore if a term is already present in an existing termbase, it could be imported.
- It will grant the interoperability of corpora, meaning that multiple corpora could be combined to cover different languages and subject fields.
- It will ease the effort made to evaluate ATE tools from both sides developers and users.

In this paper, we propose a custom form design of multi-level annotation to formalize ATE corpora in RDF format by using the OntoLex-Lemon¹ and the NLP Interchange Format - NIF² ontologies to represent termbases and corpora, respectively. Moreover, we develop a small annotated corpus to provide a proof-of-concept. The corpus and the code employed in its formalization are publicly available on GitHub³.

The remainder of this paper is organized as follows. Section 2 lays out the main feature of terms. Section 3 gives an overview of the main ATE corpora. Section 4 illustrates the proposed formalization schema. In Section 5 we describe the corpus annotation experiment. Finally, Section 6 provides conclusions.

2. Features of terms

According to ISO, a term is a “designation that represents a general concept by linguistic means” [8]. Therefore,

¹https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
³https://github.com/nicolaCirillo/lod4term
terms have linguistic and conceptual features and a sound annotation schema must account for both. The most relevant are illustrated above.

**Nested terms** A term is nested when it is contained into another (longer) term. For example, the term *competent authority of dispatch* contains both the terms *competent authority* and *dispatch*, joined by the preposition *of*.

**Discontinuous terms** A term is discontinuous when there is unrelated linguistic material between its words. Sometimes, discontinuous terms are also nested. For example, the term *prevention of pollution* is discontinuous when it appears inside the term *integrated prevention and control of pollution*.

**Term variants** A term variant is a term that expresses the same concept as other terms. For example, the terms *air pollution* and *atmospheric pollution* are term variants because they both refer to the “contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere”.

**Acronym and abbreviations** are specific kinds of term variants. Resolving abbreviations is one of the goals of the Simple Text Track at CLEF 2023.

**Terminology layer** A terminology layer is a set of terms belonging to a given subject field. For example, the terminology layer of waste management comprises terms such as *incineration plant, separate collection, and landfill*. Some ATE techniques focus on isolating terminology layers [5, 6, 9]

**Translation equivalent** A translation equivalent is a term of a natural language that denotes the same concept as another term of another natural language. For example, the term *autorità competente di spedizione* is the Italian equivalent of the English term *competent authority of dispatch*. Finding translation equivalents from comparable corpora is an ATE subtask [10].

### Table 1: Features of main ATE corpora

<table>
<thead>
<tr>
<th>corpus</th>
<th>discontinuous terms</th>
<th>nested terms</th>
<th>corpus format</th>
<th>termbase format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENIA</td>
<td>yes</td>
<td>yes</td>
<td>XML</td>
<td>none</td>
</tr>
<tr>
<td>ACL-RD TEC</td>
<td>no</td>
<td>no</td>
<td>XML; vert</td>
<td>TSV</td>
</tr>
<tr>
<td>ACTER</td>
<td>yes</td>
<td>only in the termbase</td>
<td>TSV</td>
<td>TSV</td>
</tr>
</tbody>
</table>

3. Related Work

With regards to the format of ATE benchmark corpora, there is no agreed standard. The most popular corpora are encoded in different formats as summarized in Table 1.

The GENIA corpus [11] is composed of 2000 English abstracts taken from the MEDLINE database. It is focused on the biology domain, specifically on transcription factors in human blood cells. The corpus is encoded in XML with each occurrence of a term being enclosed in the `<term>` tag. Discontinuous and nested terms are allowed. From the conceptual perspective, each term is an instance of a class defined in the GENIA ontology (e.g. the term *fibroblastic tumour* is an instance of the *Tissue* class).

Being constituted of 300 abstracts from the ACL Anthology Reference Corpus, the ACL RD-TEC corpus [12] has been developed with the intent of providing a term extraction corpus on which computational linguists are experts themselves. It is available in XML and in a vertical format (i.e. one token per line). Discontinuous and nested terms are not allowed. From the conceptual perspective terms are categorized following the guidelines (e.g. technology and method, tool and library, language resource, etc.).

The ACTER corpus [10] is composed of multiple sub-corpora covering four different subject fields and three languages. ACTER is specifically made to test ATE tools on different topics and languages while retaining a consistent annotation and, thus, comparable results. It is available in TSV (one token per line) with IOB (Inside, Outside, Beginning) or IO (Inside, Outside) tags. The list of terms found in the corpus is also made available. Discontinuous and nested terms are allowed but the latter are not represented in the IOB and IO formats. From the conceptual perspective terms are classified according to a domain-independent annotation schema composed of four labels: specific term, common term, out-of-domain term, and not term. Moreover, it distinguishes terms from Named Entities.

---


4. Proposed Format

An ATE corpus has two main components: a termbase and the actual corpus. The termbase contains the list of unique terms (types) that appear in the corpus and provides information for each of them. Conversely, the corpus contains contextualized instances (tokens) of the terms in the termbase. We propose to formalize both resources using formats based on RDF/OWL on account of their interoperability. Besides, linked data formats have already improved the benchmarking of Named Entity Recognition [13].

4.1. Termbase Representation

We propose to use the OntoLex-Lemon ontology to encode the termbase, for various reasons. First of all, OntoLex-Lemon is already a standard among terminologists [14, 15, 16]. In addition, it can represent many linguistic and conceptual information that are of interest to ATE and its subtasks.

In OntoLex-Lemon, there are three main entities: entries, concepts, and senses. Entries are instances of the LexicalEntry class. They represent linguistic units with one or more forms. For example, the term heap is an entry with a singular form (heap) and a plural form (heaps). Concepts are instances of the LexicalConcept class. They represent units of thought. For example, the concept heap is defined as "engineered facility for the deposit of solid waste on the surface". Finally, senses are instances of the LexicalSense class. They are entry-concept pairs. For example, the entry heap has multiple senses one of which couples this entry with the concept defined above. Entries are related to senses through the sense property and concepts through the lexicalizedSense property. Moreover, entries can also be directly linked to concepts through the evokes property.

Furthermore, OntoLex-Lemon allows the representation of many linguistic and conceptual features that are of interest to ATE and its subtasks (see Section 2). Term variants are easy to identify because they are entries referring to the same concept. However, OntoLex-Lemon also allows to directly link entries and senses and specifies their relation via the LexInfo ontology. Namely, the synonym property of LexInfo links two senses with the same meaning, abbreviationFor links an abbreviation to its full form, and translation links two terms that are translations of each other. OntoLex-Lemon can also represent nested terms by means of the subterm property of its decomposition module. Lastly, terminology layers can be handled by assigning senses and concepts to the respective subject field through the subject property of the Dublin Core ontology. Moreover, to grant interoperability, we propose to employ DBPedia categories [17] to represent subject fields. In this way, a term belonging to a given subject field (e.g. waste management) automatically belongs also to the subject fields it descends from (e.g. waste; sustainability and environmental management; economy and the environment), in a multi-level fashion.

4.2. Corpus Representation

To formalize the corpus, we propose to use the NIF and the POWLA\(^8\) ontologies. NIF is based on RDF/OWL and has been developed to achieve interoperability between NLP tools, language resources, and annotations. It provides multiple benefits. First of all, it does not rely strictly on tokenization like TSV formats (see Section 3). It provides support for terminology annotation and has already been used for this purpose within the FREME project [18]. The only drawback of NIF is that it cannot represent discontinuous terms. To this end, we use POWLA nodes to join NIF strings, as suggested in [19]. Then, we link each POWLA node to the corresponding LexicalSense in the termbase to produce unambiguous annotations (see Appendix A).

5. Example Corpus


5.1. Annotation

Two non-expert annotators carried out the annotation. They were instructed to identify terms in the corpus and classify them according to the subject field (i.e. law, EU law, waste management, waste management law, environment, other). Particular attention has been paid to the identification of nested and discontinuous terms (see Section 2). After the annotation phase, we asked annotators to revise the list of unique terms they found (i.e. the termbase) to delete incorrect ones and revise nested terms. Finally, we kept in the corpus only the annotations of terms that were in the revised termbases and standardized the annotation of nested terms. Namely, we removed their manual annotations and automatically tagged them according to the subterms provided in the revised termbase, thus ensuring consistency.

\(^{8}\)https://github.com/acoli-repo/powla
To estimate the inter-annotator agreement on term identification, we computed the F-score measure, similar to [12], for both the corpus and the termbase, before and after the revision process (see Table 2). Moreover, to estimate the agreement on subject fields, we computed the Fleiss’ k only on terms identified by both annotators. Inter-annotator agreement scores confirm the benefits of the revision process. Even though the agreement on the termbase shows only a little improvement after the revision (+0.034), the effect on the corpus is much more relevant (+0.149) as a result of the standardization of nested terms.

In the final dataset, we joined the annotations of both annotators and linked the resulting termbase to IATE9 by associating each concept with the respective IATE entry when it exists.

6. Conclusions and Future Work

The lack of standardization in the representation of ATE corpora constitutes a bottleneck for the evaluation of ATE tools. To address this issue, we proposed an RDF-based formalization that employs OntoLex-Lemon to represent termbases and NIF to represent corpora. We showed that these formats are able to represent the wide range of linguistic and conceptual phenomena that characterize terminology. In addition, we developed a small corpus about waste management legislation in order to provide an example of the proposed formalization.

In future, we plan to convert the major ATE corpora into the proposed format, to further improve ATE standardization. Moreover, we intend to increase the size and quality of the small ATE corpus we developed.

Acknowledgments

The authors contributed to this paper as follows. Sections 1, 2, and 6 are attributed to Daniela Vellutino while sections 3, 4, and 5 are attributed to Nicola Cirillo.

References

[12] B. Qasemizadeh, A.-K. Schumann, The acl rd-

---

Table 2: Inter-annotator agreement

<table>
<thead>
<tr>
<th></th>
<th>termbase (F-score)</th>
<th>corpus (F-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>before revision</td>
<td>0.440</td>
<td>0.470</td>
</tr>
<tr>
<td>after revision</td>
<td>0.474</td>
<td>0.619</td>
</tr>
<tr>
<td>subject fields</td>
<td>0.707</td>
<td></td>
</tr>
</tbody>
</table>

---

https://iate.europa.eu/home
tec 2.0: A language resource for evaluating term extraction and entity recognition methods, European Language Resources Association (ELRA), 2016, pp. 1862–1868. URL: https://aclanthology.org/L16-1294.


A. Example of RDF files

@prefix dbc: <https://dbpedia.org/page/Category> .
@prefix dct: <http://purl.org/dc/terms> .
@prefix decomp: <http://www.w3.org/ns/lemon/decomp#> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix ontlex: <http://www.w3.org/ns/lemon/ontlex#> .
@prefix termbase: <http://example.com/termbase/> .

termbase:entry_rifiuto_delle_industrie_estrattive a ontlex:MultiwordExpression ;
    decomp:subterm termbase:entry_industria_estrattiva ,
    termbase:entry_rifiuto ;
    ontlex:canonicalForm termbase:form_rifiuto_delle_industrie_estrattive ;
    ontlex:otherForm termbase:form_rifiuti_delle_industrie_estrattive ;
    ontlex:sense termbase:rifiuto_delle_industrie_estrattive_sense1

termbase:form_rifiuto_delle_industrie_estrattive a ontlex:Form ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:number lexinfo:singular ;
    ontlex:writtenRep "rifiuto delle industrie estrattive"@it .

termbase:form_rifiuti_delle_industrie_estrattive a ontlex:Form ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:number lexinfo:plural ;
    ontlex:writtenRep "rifiuti delle industrie estrattive"@it .

termbase:rifiuto_delle_industrie_estrattive_sense1 a ontlex:LexicalSense ;
    ontlex:isLexicalizedSenseOf termbase:concept_rifiuto_delle_industrie_estrattive ;
    ontlex:isSenseOf termbase:entry_rifiuto_delle_industrie_estrattive ;
    dct:subject dbc:Waste_management ;
    lexinfo:synonym termbase:rifiuto_derivante_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_generato_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_prodotto_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_proveniente_dalle_industrie_estrattive_sense1 .

termbase:concept_rifiuto_delle_industrie_estrattive a ontlex:LexicalConcept ;
    dct:subject dbc:Waste_management ;
    ontlex:lexicalizedSense termbase:rifiuto_delle_industrie_estrattive_sense1 ,
    termbase:rifiuto_derivante_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_generato_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_prodotto_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_proveniente_dalle_industrie_estrattive_sense1 ;
    ontlex:isEvokedBy termbase:entry_rifiuto_delle_industrie_estrattive_sense1 ,
    termbase:entry_rifiuto_derivante_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_generato_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_prodotto_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_proveniente_dalle_industrie_estrattive .

Figure 1: Example of the termbase.
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix powla: <http://purl.org/powla/powla.owl#> .
@prefix termbase: <http://example.com/termbase/> .
@prefix corpus: <http://example.com/corpus/> .

_:terms a powla:Root .
_:term1 a powla:Node;
  powla:string "rifìuti";
  powla:hasParent _:terms;
  itsrdf:term "yes";
  itsrdf:termInfoRef termbase:rifiuto_sense1 .

_:term2 a powla:Node;
  powla:string "industrie estrattive";
  powla:hasParent _:terms;
  itsrdf:term "yes";
  itsrdf:termInfoRef termbase:industria_estrattiva_sense1 .

_:term3 a powla:Node;
  powla:string "rifiuti delle industrie estrattive";
  powla:hasParent _:terms;
  itsrdf:term "yes";
  itsrdf:termInfoRef termbase:rifiuto_delle_industrie_estrattiva_sense1 .

corpus:doc1 a nif:Context;
  nif:OffsetBasedString;
  nif:isString "Direttiva 2006/21/CE del Parlamento europeo e del Consiglio ... " .

corpus:doc1#offset_105_112 a nif:OffsetBasedString;
  nif:Word;
  powla:Node;
  nif:anchorOf "rifìuti";
  nif:referenceContext corpus:doc1;
  powla:hasParent _:term1,
  _:term3 .

corpus:doc1#offset_113_118 a nif:OffsetBasedString;
  nif:Word;
  powla:Node;
  nif:anchorOf "delle";
  nif:referenceContext corpus:doc1;
  powla:hasParent _:term3;
  powla:next corpus:doc1#offset_113_118 .

corpus:doc1#offset_113_118 a nif:OffsetBasedString;
  nif:Word;
  powla:Node;
  nif:anchorOf "industrie";
  nif:referenceContext corpus:doc1;
  powla:hasParent _:term2,
  _:term3;
  powla:next corpus:doc1#offset_129_139 .

corpus:doc1#offset_129_139 a nif:OffsetBasedString;
  nif:Word;
  powla:Node;
  nif:anchorOf "estrattive";
  nif:referenceContext corpus:doc1;
  powla:hasParent _:term2,
  _:term3 .

Figure 2: Example of the corpus