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

Nicola Cirillo[1], Daniela Vellutino[1]

[1]University 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.

Italiano. I corpora principali impiegati nella valutazione degli algoritmi di Estrazione Automatica di Termini sono codificati in formati diversi. Purtroppo, nessuno di questi formati permette di rappresentare l'ampia gamma di fenomeni linguistici legati alla terminologia. Per affrontare la questione, proponiamo di codificare i corpora di Estrazione Automatica di Termini in RDF usando le ontologie OntoLex-Lemon e NLP Interchange Format. Inoltre, abbiamo sviluppato un piccolo corpus italiano riguardante la legislazione della gestione dei rifiuti per fornire un esempio della formalizzazione proposta.

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-Lemon[1] and the NLP Interchange Format - NIF[2] 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 GitHub[3]

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,

[1]https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

[2]https://nif.readthedocs.io/en/latest/

[3]https://github.com/nicolaCirillo/lod4term

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

nicirillo@unisa.it (N. Cirillo); dvellutino@unisa.it (D. Vellutino)

0000-0002-2107-1313 (N. Cirillo); 0000-0002-2525-7940 (D. Vellutino)

corpus	discontinuous terms	nested terms	corpus format	termbase format
GENIA	yes	yes	XML	none
ACL-RD TEC	no	no	XML; vert	TSV
ACTER	yes	only in the termbase	TSV	TSV

Table 1
Features of main ATE corpora

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".[4] **Acronym** and **abbreviations** are specific kinds of term variants. Resolving abbreviations is one of the goals of the Simple Text Track at CLEF 2023.[5]

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].

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 subcorpora 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] https://iate.europa.eu/entry/result/3567909/en
[5] http://simpletext-project.com/2023/clef/tasks

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 are 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".[6] 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 `lexicalicalizedSense` 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[7] 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

In order to test our approach and provide a proof-of-concept, we run an annotation experiment on a European directive, namely the Italian version of the *Directive 2006/21/EC of the European Parliament and of the Council of 15 March 2006 on the management of waste from extractive industries and amending Directive 2004/35/EC* (26,882 tokens).

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.

[6] https://iate.europa.eu/entry/result/3504812/en
[7] https://lexinfo.net/
[8] https://github.com/acoli-repo/powla

	termbase (F-score)	corpus (F-score)
before revision	0.440	0.470
after revision	0.474	0.619
subject fields (Fleiss' k)	0.707	

Table 2
Inter-annotator agreement

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 IATE[9] 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.

[9] https://iate.europa.eu/home

References

[1] D. Vellutino, R. Maslias, F. Rossi, Verso l'interoperabilità semantica di iate. studio preliminare per il dominio "gestione dei rifiuti urbani", Terminologie specialistiche e diffusione dei saperi (2016) 1–240.

[2] D. Vellutino, R. Maslias, F. Rossi, C. Mangiacapre, M. P. Montoro, Verso l'interoperabilità semantica di iate. studio preliminare sul lessico dei fondi strutturali e d'investimento europei, Diversité et Identité Culturelle en Europe/Diversitate si Identitate Culturala in Europa (2016) 1–254.

[3] D. Vellutino, L'italiano istituzionale per la comunicazione pubblica, il Mulino, 2018.

[4] A. Lenci, S. Montemagni, V. Pirrelli, G. Venturi, Ontology learning from italian legal texts, in: Law, Ontologies and the Semantic Web, IOS Press, 2009, pp. 75–94.

[5] F. Bonin, F. Dell'Orletta, S. Montemagni, G. Venturi, A contrastive approach to multi-word extraction from domain-specific corpora, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/553_Paper.pdf.

[6] P. Drouin, M.-C. L'Homme, B. Robichaud, Lexical profiling of environmental corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[7] ISO, ISO 30042:2019 - Management of terminology resources - TermBase eXchange (TBX), International Organization for Standardization, 2019.

[8] ISO, ISO 1087:2019 - Terminology work and terminology science, International Organization for Standardization, 2019.

[9] N. Cirillo, Isolating terminology layers in complex linguistic environments: a study about waste management (short paper), in: Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023), volume 3427, CEUR Workshop Proceedings, 2023. URL: https://ceur-ws.org/Vol-3427/short3.pdf.

[10] A. Rigouts Terryn, V. Hoste, E. Lefever, In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora, Language Resources and Evaluation 54 (2020) 385–418. doi:10.1007/s10579-019-09453-9.

[11] J. D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus - a semantically annotated corpus for bio-textmining, Bioinformatics 19 (2003). doi:10.1093/bioinformatics/btg1023.

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

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.

[13] M. Röder, R. Usbeck, A.-C. N. Ngomo, Gerbil – benchmarking named entity recognition and linking consistently, Semantic Web 9 (2018) 605–625. doi:10.3233/SW-170286.

[14] P. Martín-Chozas, T. Declerck, Representing multilingual terminologies with ontolex-lemon, in: Proceedings of the 1st International Conference on Multilingual Digital Terminology Today (MDTT 2022), CEUR Workshop Proceedings, 2022.

[15] S. Piccini, F. Vezzani, A. Bellandi, Tbx and 'lemon': What perspectives in terminology?, Digital Scholarship in the Humanities 38 (2023) i61–i72. URL: https://doi.org/10.1093/llc/fqad025. doi:10.1093/llc/fqad025.

[16] M. Fiorelli, A. Stellato, T. Lorenzetti, A. Turbati, P. Schmitz, E. Francesconi, N. Hajlaoui, B. Batouche, Towards ontolex-lemon editing in vocbench 3, AIDAinformazioni, Rivista di scienze dell'informazione (2018).

[17] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. V. Kleef, S. Auer, C. Bizer, Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2015) 167–195. doi:10.3233/SW-140134.

[18] M. Dojchinovski, F. Sasaki, T. Gornostaja, S. Hellmann, E. Mannens, F. Salliau, M. Osella, P. Ritchie, G. Stoitsis, K. Koidl, M. Ackermann, N. Chakraborty, Freme: Multilingual semantic enrichment with linked data and language technologies, volume 8, European Language Resources Association (ELRA), 2016, pp. 4180–4183. URL: https://aclanthology.org/L16-1660.

[19] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data Representation, Generation and Applications, 1 ed., Springer Cham, 2020. doi:10.1007/978-3-030-30225-2.

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 ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix termbase: <http://example.com/termbase/> .

termbase:entry_rifiuto_delle_industrie_estrattive a ontolex:MultiwordExpression ;
  decomp:subterm termbase:entry_industria_estrattiva ,
    termbase:entry_rifiuto ;
  ontolex:canonicalForm termbase:form_rifiuto_delle_industrie_estrattive ;
  ontolex:otherForm termbase:form_rifiuti_delle_industrie_estrattive ;
  ontolex:sense termbase:rifiuto_delle_industrie_estrattive_sense1
  ontolex:evokes termbase:concept_rifiuto_delle_industrie_estrattive .

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

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

termbase:rifiuto_delle_industrie_estrattive_sense1 a ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf termbase:concept_rifiuto_delle_industrie_estrattive ;
  ontolex: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 ontolex:LexicalConcept ;
  dct:subject dbc:Waste_management ;
  ontolex: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 ;
  ontolex: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 xsd: <http://www.w3.org/2001/XMLSchema#> .
@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 "rifiuti" ;
  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_estrattive_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 "rifiuti" ;
  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