Ureña Gómez-Moreno, Pedro (2018) Learning IS-A relations from specialized-domain texts with co-occurrence measures. Journal of Computer-Assisted Linguistic Research 2: 21-38.

Learning IS-A relations from specialized-domain texts with co-occurrence measures

Pedro Ureña Gómez-Moreno
University of Granada, Spain
Corresponding author: pedrou@ugr.es

Received: 3 February 2018 / Accepted: 5 May 2018 / Published: 12 July 2018

Abstract
Ontology enrichment is a classification problem in which an algorithm categorizes an input conceptual unit in the corresponding node in a target ontology. Conceptual enrichment is of great importance both to Knowledge Engineering and Natural Language Processing, because it helps maximize the efficacy of intelligent systems, making them more adaptable to scenarios where information is produced by means of language. Following previous research on distributional semantics, this paper presents a case study of ontology enrichment using a feature-extraction method which relies on collocational information from corpora. The major advantage of this method is that it can help locate an input unit within its corresponding superordinate node in a taxonomy using a relatively small number of lexical features. In order to evaluate the proposed framework, this paper presents an experiment consisting of the automatic classification of a chemical substance in a taxonomy of toxicology.

Keywords: ontology learning, ontology enrichment, taxonomy, corpus linguistics, co-occurrence

1. Introduction

Ontology learning is an area of study within Knowledge Engineering which aims to create and/or expand conceptual ontologies using automatic methods or minimum human intervention. Ontologies are computer-readable schemas representing abstract units of thought, such as ANIMAL, HUMAN, OBJECT, etc., which in turn represent conceptualizations of lexical units from natural languages, such as *animal*, in English and Spanish, or *Tier* in German. The most basic types of ontologies are taxonomies which organize concepts according to parent-child relationships, where parents stand for prototypical concepts of generic reference (e.g. ANIMAL), whilst children are subordinate concepts whose specific meaning is not shared by the rest of children (e.g. DOG, CAT, COW). The term "learning" has here a computational meaning, namely the acquisition of conceptual information from different types of data sources,

DOI: https://doi.org/10.4995/jclr.2018.9916
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

such as text corpora, SQL databases or XML files, along with the networks of relationships that are established among groups of concepts. Ontology learning has been extensively studied in the literature due to its importance in the development and/or enhancement of many applications, including decision-making systems, information-retrieval algorithms or the semantic web.

There are two main subtasks of ontology learning that must be distinguished, namely construction and enrichment. Ontology construction involves the induction from scratch of a complete ontology, typically by first retrieving keywords from one or more data sources like the ones mentioned above, then clustering them into concepts and, finally, identifying the underlying relationships that take place among them (Cimiano 2006; Buitelaar and Cimiano 2008). Ontological relationships can be of two types: taxonomic (i.e. is-a relationships) or non-taxonomic (i.e. meronymy, agent-object, etc.). On the other hand, the goal of ontology enrichment, which is, a priori, a much less demanding task than the former, is to develop algorithms for classifying new concepts under their corresponding host superordinates in an existing ontology (Faatz and Steinmetz 2003, 2005). Nevertheless, ontology enrichment becomes complex in the case of highly fine-grained taxonomies containing hundreds or even thousands of concepts, hence the need to develop computer-assisted methodologies that allow the categorization of new concepts with minimum time investment and computational cost.

This paper presents a case study of ontological enrichment using a feature-based method which relies on statistical information from corpus data. The method, which will be explained in depth in Section 3, focuses on the process of obtaining relevant collocates of a lexical unit from unstructured sources, i.e. texts in natural language, and the exploitation of such features for the automatic classification of the unit in a target taxonomy. The statistical analysis of co-occurrence patterns applied to fields such as concept learning or document classification is not new to the literature. Prior research including Agirre, Alfonseca and López de Lacalle (2004), Alfonseca and Manandhar (2002), Fotzo and Gallinari (2004), and De Knijff, Frasincar and Hogenboom (2013) are among the first authors to study distributional techniques for learning tasks, although they are mainly focused on the expansion of general-domain lexical repositories like WordNet (Princeton University, 2010). The goal of the present paper is to present further evidence on the strengths of the feature-based approach and confirm the results obtained in a previous study in which the method was applied for learning taxonomic relationships in a corpus of virology (Ureña and Mestre-Mestre 2017). More specifically, in Ureña and Mestre 2017 we showed that the application of a Normalized Pointwise Mutual Information metric on a list of collocates of the target domain was a successful method to classify a term within the corresponding taxonomy. The results obtained in the experiment in the present paper point in the same direction, while we also include a more detailed discussion on the adequacy of other statistical tests for the classification task.

The method can be applied to enrich any terminological-conceptual taxonomy. In this paper, nevertheless, it will be proposed for the enrichment of FunGramKB's ontology model. As explained elsewhere in the literature (Periñán-Pascual and Arcas Túnez 2010; Periñán-Pascual and Mairal Usón 2010), FunGramKB is a knowledge base that is strongly grounded both in the semantic representation of conceptual constructs that replicate human knowledge, and in the

formalization of linguistic information, which represents grammatical, constructional and lexical data from diverse languages. The proposed method seeks to enhance the informativeness of the conceptual module of the knowledge base and, more specifically, the specialized repositories which are envisaged to enlarge the conceptual engine. It is important to notice that FunGramKB has two separate components for ontological and lexical information, respectively; however, for the sake of simplicity, in this paper linguistic units (e.g. *animal*) will be considered as equivalent to conceptual units (e.g. ANIMAL), and the expressions "term", "keyword" and "concept" will be used interchangeably.

This paper is organized as follows. This introductory section is followed by a brief review of the literature on ontology learning in Section 2. Section 3 presents the method for the acquisition of taxonomic relationships and explains the major steps of an experiment carried out to evaluate the method. Section 4 presents the results of the experiment. Finally, Section 5 summarizes the main findings and conclusions.

2. Related Work

Knowledge engineering has been studied for some years now and the literature in this area of research is extensive. Ever since the publication of the first influential works on ontologies (Gruber 1995; Maedche and Staab 2001) and the later appearance of various reference studies (Cimiano 2006; Buitelaar and Cimiano 2008), there has been a growing interest in developing new techniques for the automatic extraction of terms and concepts from texts. The study of ontology learning in particular has gained momentum, which is evidenced by the publication of several state-of-the-art works reviewing the advances in the field (Shamsfard and Barforoush 2003; Gómez-Pérez and Manzano-Macho 2004; Biemann 2005; Zouaq and Nkambou 2010; Petasis et al. 2011; Hazman, El-Beltagy and Rafea 2011; Clark et al. 2012; Wong, Liu and Bennamoun 2012; Gherasim et al. 2013; Lehmann and Völker 2014).

There are two major approaches to learning from unstructured sources that are clearly identified in the literature. The linguistic approach involves the use of pre-established lexico-syntactic structures, such as "NP$_X$ *and other* NP$_Y$" for predicting the occurrence of, for instance, hypernyms-hyponym relationships in raw texts (Hearst 1992). This method has been proved to attain high precision but lower recall rates. Similarly, the semantic approach involves the combination of Hearst-like patterns with lexico-semantic patterns. The major advantage over the former is that this proposal shows improved recall rates of the target expressions, since the lexical slots have greater generalizing power (IJntema et al. 2012). On the other hand, the statistical approach encompasses a range of techniques, including semantic similarity measures and probabilistic modeling of linguistic data (for an overview of methods, see Meijer, Frasincar and Hogenboom 2014). The major advantage of this approach is that it overcomes the language-specificity limitation and has also paved the way to the application of machine learning techniques.

Progress has also been made at a more practical level with the implementation of a number of algorithms and software applications which, based on the approaches mentioned above, enable the creation and/or enrichment of ontologies. Especially relevant to the present study are

computerized solutions for learning concepts from specialized texts, such as OntoLearn Reloaded, which proposes an algorithm for constructing ontologies from scratch by exploiting dictionary-like definitions using graph theory (Velardi, Faralli and Navigli, 2013). Another well-known example of software is Text2Onto, which uses the skewed divergence similarity measure for the assignment of new instances to specific concepts in a reference ontology. Text2Onto represents instances using meta-level primitives which allow simplified exportation of learned structures to other ontologies (Cimiano and Völker, 2005).

3. Method

This section describes the Ontology Enrichment Method (henceforth OEM) aimed at the learning of taxonomic relations from domain text collections. The general workflow of the OEM involves selecting an input term, which we aim to classify into a pre-existing reference taxonomy, and automatically assigning it to one superordinate unit at the lowermost layer in the taxonomy. For this classification task, the OEM follows a feature-matching process between the input and the superordinates. The main rationale, therefore, is that the superordinate sharing the highest number of relevant features with the input term will be considered the strongest candidate to subsume the latter. In order to evaluate the method, we conducted an experiment using English as a case study and whose aim was to categorize a chemical substance into a target taxonomy of toxic compounds.

Before the experiment is presented it is noteworthy to mention that the OEM adopts the so-called bag-of-words model of text representation, in which linguistic tokens are processed as independent strings of text, i.e. without any assumption about their order in the sentence or about any dependency relationships, either paradigmatic (i.e. collocations and colligations), or syntagmatic (e.g. polysemy or homonymy). As has been recursively claimed in the literature, even though this model makes a naïve assumption on the nature of language, which is primarily a system of dependencies, it yields accurate results in most tasks of language processing. In addition to the bag-of-words model, it is important to mention that the OEM is grounded on the distributional hypothesis of language, which postulates that words with similar meanings share similar contexts, i.e. similar co-occurrence patterns (Harris 1954).

The method relies on three major data sources, namely a reference taxonomy, an encyclopedic database and a domain corpus. We will utilize the reference taxonomy both to select a group of units that we intend to classify automatically and, as explained in Section 3.1, to validate the results of the experiment. On the other hand, the domain corpus will be used to extract the features of the input term, while the encyclopedic database will be used to extract the features of the superordinates. Using the corpus as the data source for the input keyword will help to simulate real learning contexts where the OEM must classify new expressions from unstructured sources, such as websites, Wikipedia articles or scientific texts online. The role of the three data types in the enrichment task are explained in more depth below.

3.1 Data sources

3.1.1 The reference taxonomy

The first step in the OEM is to select an existing taxonomy of a specialized domain, such as biology, medicine or electronics. As mentioned above, the taxonomy will constitute the gold standard, i.e. it will be used to evaluate whether the OEM classifies the input in the correct node shown in the standard. For the experiment we chose the lexical taxonomy provided by the Agency for Toxic Substances and Disease Registry (ATSDR) which contains 14 major classes of toxic chemical compounds, each in turn containing various instance substances.[1] Starting from a selection of 6 classes, we chose *ethylbenzene* as the input term, and *pyridine*, *toluene*, *benzene*, *ammonia*, *chlorine*, *phenol* and *iodine* as the potential superordinates. In addition to the term *ethylbenzene*, two other units — instead of one — were selected from Class 2, namely *toluene* and *benzene*, based on the hypothesis that both these units should have more features in common with *ethylbenzene*. Table 1 summarizes the set of instances for the experiment.

Class	**Class name**	**Instance**
Class 1	benzidines / aromatic amines	*pyridine*
Class 2	hydrocarbons	*toluene, benzene, ethylbenzene*
Class 3	inorganic substances	*ammonia*
Class 4	metals	*chlorine*
Class 5	phenols / phenoxy acids	*phenol*
Class 6	radionuclides	*iodine*

TABLE 1. CHEMICAL SUBSTANCES CHOSEN FOR THE EXPERIMENT

The selection shown above was carried out randomly, with no previous estimation of the lexical features of each unit and without considering potential membership affinities across classes. The only criterion that we followed, nevertheless, was that the input unit showed a representative occurrence in the corpus, so that the experiment explained in the sections below could be viable.

[1] https://www.atsdr.cdc.gov/substances/ToxChemicalClasses.asp

3.1.2 The encyclopedic database

In addition to the taxonomy, the ATSDR is important for the present study, because it contains toxicological profiles for every substance included in the repository, which can be used to mine lexical features of the superordinates. The profiles show detailed information on aspects ranging from preventive measures, environmental hazards, health effects on humans, the chemical and physical structures, as well as other aspects about production, import, disposal, etc. For the experiment we focused on the subsection within the profile which deals with public health and which encompasses information on the following aspects (see also Annex I):

a. What is this substance?

b. What happens to it when it enters the environment?

c. How might I be exposed to it?

d. How can it enter and leave my body?

e. How can it affect my health?

f. Is there a medical test to determine whether I have been exposed to it?

g. What recommendations has the federal government made to protect human health?

h. Where can I get more information?

To obtain features of the superordinate from the toxicological profiles, we preprocessed each of them, by first tokenizing (i.e. obtaining separate word units from strings of text) and then lemmatizing them (i.e. reducing the lexical units to their base form by removing derivational or inflectional affixes) and carrying out stopword removal (i.e. filtering high-occurrence functional units, such as pronouns or conjunctions). For text processing, both at this and later stages in the experiment, we used Data Mining Encountered (DAMIEN) (Periñán-Pascual 2017), a robust online toolkit which enables the completion of various language-related tasks including corpus processing, statistical analysis, data mining and evaluation.[2]

3.1.3 The domain corpus

PLOS ONE is a scholarly journal that contains a large collection of scientific papers on research areas such as ecology, physical sciences, earth sciences, information sciences, among many others,[3] and here it will be used as an ad-hoc untagged corpus for mining collocates of the input term. In order to access PLOS ONE, we used AntCorGen, an application specifically designed for compiling texts from the journal database (Anthony 2014).[4] For the experiment we obtained a sample collection by running a query of *ethylbenzene* based on three main user-defined parameters. First, we set a retrieval threshold of 200 hits (i.e. maximum number of documents containing the input term) and, as a result, we obtained 69 documents out of which we retrieved 184 concordances of *ethylbenzene*. Below are some example concordances of *ethylbenzene*:

[2] http://www.fungramkb.com/nlp/aspx

[3] https://www.plos.org/. Last accessed March 2018.

[4] http://www.laurenceanthony.net/software/antcorgen/

(1) accounted for in the model, a rough average of these stoichiometries (1.47:1) based on toluene, *ethylbenzene*, xylene, and hexadecane degradation [18] was used to determine the amount of bicarbonate per mole of carbon. (_10_1371_journal_pbio_0030077.txt)

(2) PBPK models that could act as a starting template to build a new model, *ethylbenzene* was used as a case study. Six chemicals with varying structural similarities towards. (_10_1371_journal_pcbi_1004495.txt)

(3) each series. Internal standards consisted of a solution containing 100 ng/μl of benzene, toluene, *ethylbenzene* and xylenes in methanol [16], [17], [18]. Standard Tenax tubes were prepared by. (_10_1371_journal_pone_0013423.txt)

Second, we restricted the query to the body of the articles in PLOS ONE, thus excluding other sections, such as the abstract or the acknowledgments. Third, we further restricted the query by means of the field filter in AntCorGen, which allows to choose between one or more subareas of study in the scope of the journal. One major advantage of this filter is that it helps overcome potential cross-field ambiguities that may appear in relation to the term(s) under study. In our case, we restricted the query to the field of toxicology, which ensured that the sense of the units involved in the experiment was essentially related to this domain.

3.2 Statistical component

Once the corpus sample was collected, we proceeded to its analysis with the purpose of obtaining a set of relevant features of the input term. With this aim we performed the following steps using DAMIEN. First, we extracted the concordances of the input term from the sample and then retrieved a list of lemmatized unigrams from it (i.e. one-word expressions). Second, we submitted the lemmatized set to a stopword-list filtering process. Third, the set was processed statistically to find a list of collocates of the input term, i.e. expressions which appear in the same context of the input with statistically significant frequency. For this purpose, we applied three different metrics: (a) the Pointwise Mutual Information score, whose formula is indicated below (Fano 1961, mentioned in Jurafsky and Martin 2008):

$$PMI(w,x) = log_2 \frac{P(w,c)}{P(w)P(c)}$$

(b) the Pearson chi-square test, whose formula is shown below:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where *O* stands for the observed frequency in the corpus of both lexical units being compared and *E* stands for their expected frequency, and (c) the log-likelihood ratio (Cressie and Read 1989):

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

Fourth, we established various ranges within the list of collocates of the input term along with their corresponding weight relative to higher or lower scores. The ranges were established manually based on the observation of clear-cut groupings of units around specific scores. The ranges and their associated weights are shown in Table 2.

Range	Weight	PMI	Chi-square	Log-likelihood
1	0.9	0.0317	45.2213 > x > 34.5825	14.2141 > x > 7.4771
2	0.8	0.0316	34.5824 > x > 13.9350	7.4770 > x > 6.0151
3	0.7	0.0315	13.9349 > x > 7.6887	6.0150 > x > 3.8692
4	0.6	0.0314	7.6886 > x > 3.5593	3.8691 > x > 0.9620
5	0.5	0.0313	3.5592 > x > 0.7483	0.9619 > x > 0.2938
6	0.4	0.0312	0.7482 > x > 0.0890	0.2937 > x > 0.2294
7	0.3	0.0311 > x > 0.0310	0.0889 > x > 0.0431	0.2293 > x > 0.1024
8	0.2	0.0309 > x > 0.0303	0.0430 > x > 0.0205	0.1023 > x > 0.0403
9	0.1	0.0302 > x > 0.0105	0.0204 > x > 0	0.0404 > x > 0

TABLE 2. RANGE AND WEIGHTS IN THE LIST OF COLLOCATES OF THE INPUT TERM

Fifth, we submitted the list of features, both of the input term and the superordinate terms, to a matching process. As a result, every superordinate feature that did not match the same feature in the list of collocates was removed from the experimental dataset. On the other hand, the matching features of each superordinate were assigned a corresponding weight based on the schema shown in Table 2. Finally, we compared the weights across superordinates by using the standard z score (see Section 4). The matching-scoring system along these lines was aimed at

revealing which superordinate showed greater weight and was thus more distributionally related to *ethylbenzene*. Figure 1 summarizes the main steps described in Section 3.

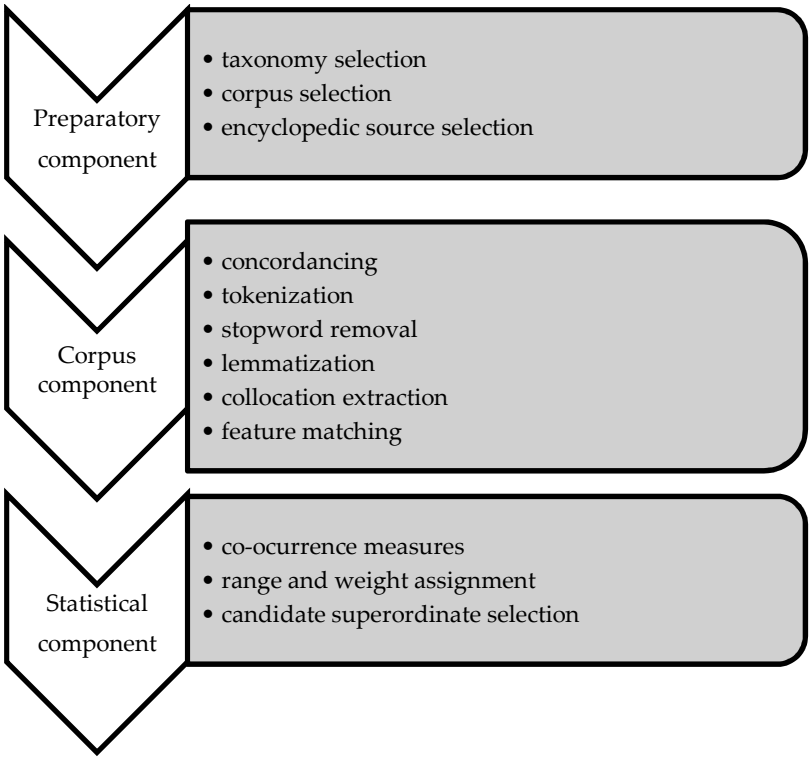

FIGURE 1. THE MAIN STEPS IN THE OEM

The next section describes the main results obtained from the application of the OEM on the corpus sample of toxic substances.

4. RESULTS

The query of the input term *ethylbenzene* in the PLOS journals resulted in 69 documents, totaling 184 concordances. Upon this sample, the statistical analysis of the concordances was carried out which resulted in a set of 888 collocates in the case of the PMI, 940 collocates in the case of chi-square and 938 collocates in the case of the log-likelihood. Table 3 summarizes the topmost collocates of the input (*ethylbenzene* has been excluded from the three collocate lists).

Collocates	PMI	Collocates	Chi-square	Collocates	Log-likelihood
xylene	0.0317	*chlorinate*	45.2213	*inhalation*	14.2140
styrene	0.0316	*coeff*	45.2213	*chlorinate*	14.2140
study	0.0316	*context*	45.2213	*coeff*	14.2140

chemical	0.0316	*dibenzofuran*	45.2213	*context*	14.2140
compound	0.0316	*genase*	45.2213	*dibenzofuran*	14.2140
pbpk	0.0316	*gradient*	45.2213	*genase*	14.2140
model	0.0316	*hepatocyte*	45.2213	*gradient*	14.2140
aromatic	0.0316	*liver*	45.2213	*hepatocyte*	14.2140
hydrocarbon	0.0316	*mea*	45.2213	*liver*	14.2140
case	0.0316	*observe*	45.2213	*mea*	14.2140

TABLE 3. TOP-SCORING COLLOCATES OF THE INPUT TERM

As described in Section 3, each vector of features corresponding to every potential superordinate was compared to the vector of features corresponding to the input term via feature-matching. The results of this process are summarized in Table 4.

Superordinates	**Subordinates**	**Shared features with the input term**	**Weight of the shared features**					
			Calculated based on the collocates' PMI of the input term		**Calculated based on the collocates' Chi-square of the input term**		**Calculated based on the collocates' Log-likelihood of the input term**	
			Total	**Z**	**Total**	**Z**	**Total**	**Z**
inorganic substances	*ammonia*	112	38.6	0.48	29.5	0.24	31	0.23
hydrocarbons	*benzene*	133	45.4	**1.40**	33.9	**0.97**	35.7	**0.99**
metals	*chlorine*	87	32	-0.40	28.6	0.09	30.2	0.10

radionuclides	*iodine*	97	33.2	-0.24	25.1	-0.49	26.6	-0.48
phenols / phenoxy acids	*phenol*	82	28.9	-0.82	24.6	-0.57	25.9	-0.59
benzidines / aromatic amines	*pyridine*	63	24.5	-1.41	18.3	-1.61	19.7	-1.59
hydrocarbons	*toluene*	120	42.3	**0.98**	36.2	**1.35**	38	**1.36**

TABLE 4. SHARED FEATURES BETWEEN THE INPUT AND THE SUPERORDINATES

The results in Table 4 suggest that the two instances showing both a greater number of shared features with *ethylbenzene* and higher aggregate weights are *toluene* and *benzene*. These distributional similarities point to the class-membership relation of the three units as shown in the taxonomy in Table 1, which is consistent with the research goals set at the introduction of the study. It must be also highlighted that the three statistical measures selected for the experiment have offered similar results, which reinforces the importance of co-occurrence measures in taxonomical classification tasks.

5. Conclusions

Ontology learning is of great importance to several academic disciplines including Natural Language Processing and Knowledge Engineering, because it provides a rapid and reliable means for developing both general-domain and specialized ontologies. Ultimately, creating new ontologies and/or expanding pre-existing ones offers opportunities at different levels, for example, in the development of computational systems that cope with the data-deluge problem originating in the era of the big data. Another example is the creation of intelligent agents that can acquire knowledge autonomously, while linking this knowledge to other data structures already present in them. Nonetheless, creating ontologies requires a great deal of material and human resources, and thus even greater efforts are needed to streamline the process of acquiring, modeling and classifying new concepts.

This paper has presented a stepwise methodology for classifying lexical and/or conceptual units automatically in the corresponding node within a target taxonomy. The method, which elaborates on previous linguistic approaches to conceptual acquisition found in the literature, is both language- and domain-independent, so that it can be used in varied learning contexts irrespective of the source language. Our main contribution has been to reinforce the importance of distributional models based on feature-extraction as an effective technique for the enrichment of ontologies. The general principle of the method is that the taxonomic relationships of an input term can be learned by first exploiting co-occurrence information from a corpus and then matching this information to the set of features that represent the superordinate concepts in a target ontology. To evaluate the performance of the method, the paper presented an experiment

carried out in the area of toxicology in which one input chemical compound was extracted from a pre-existing gold taxonomy with the aim of classifying it automatically into the node of origin. The method was presented as proof-of-concept and would therefore require future implementation for practical use within real computational environments.

Acknowledgements

This article is based on research carried out within the framework of the Project FFI2014-53788-C3-1-P, which is funded by the Spanish Ministry of Economy and Competitiveness, and entitled: Development of a virtual laboratory for natural language processing from a functional paradigm.

References

Agirre, Eneko, Alfonseca, Enrique, and Oier López de Lacalle. 2004. Approximating Hierarchy-based Similarity for WordNet Nominal Synsets Using Topic Signatures. Proceedings of GWC-04, 2nd global WordNet Conference, edited by Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, and Piek Vossen, 15–22. The Global Wordnet Association.

Alfonseca, Enrique, and Suresh Manandhar. 2002. "Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures." In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. EKAW 2002*. *Lecture Notes in Computer Science*, edited by Asunción Gómez-Pérez, and Richard Benjamins, Vol. 2473, 1–7. Berlin, Heidelberg: Springer. doi: 10.10073-540-45810-7.

Anthony, Laurence. 2018. AntCorGen (Version 1.1.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Biemann, Chris. 2005. "Ontology learning from text: A survey of methods." *LDV forum* 20(2): 75–93.

Buitelaar, Paul, and Philipp Cimiano, eds. 2008. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Amsterdam: IOS Press.

Cimiano, Philipp. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Berlin, Heidelberg: Springer.

Cimiano Philipp, and Johanna Völker. 2005. "Text2Onto." In *Natural Language Processing and Information Systems*, edited by Andrés Montoyo, Rafael Muñoz R, and Elisabeth Métais, 227-238. Lecture Notes in Computer Science, vol 3513. Berlin, Heidelberg: Springer. doi.org/10.1007/11428817_21.

Clark, Malcolm, Kim, Yunhyong, Kruschwitz, Udo, Song, Dawei, Albakour, Dyaa, Dignum, Stephen, Cerviño Baresi, Ulises, Fasli, Maria, and Anne De Roeck. 2012. "Automatically Structuring Domain Knowledge from Text: An Overview of Current Research." *Information Processing and Management* 48(3): 552–568. doi: dx.doi.org/10.1016/j.ipm.2011.07.002.

Cressie, Noel and Timothy R. C. Read. 1989. "Pearson's X2 and the Loglikelihood Ratio Statistic G2: A comparative review." *International Statistical Review* 57(1): 19–43.

De Knijff, Jeroen, Frasincar, Flavius, and Frederik Hogenboom. 2013. "Domain Taxonomy Learning from Text: The Subsumption Method versus Hierarchical Clustering." *Data and Knowledge Engineering* 83: 54–69. doi: dx.doi.org/10.1016/j.datak.2012.10.002.

Faatz, Andreas, and Ralf Steinmetz. 2003. "Ontology Enrichment with Texts from the WWW." In Proceedings of the 2nd ECML/PKDD Semantic Web Mining Workshop.

Faatz, Andreas, and Ralf Steinmetz. 2005. "An Evaluation Framework for Ontology Enrichment." In *Ontology Learning from Text: Methods, Applications and Evaluation*, edited by Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, number 123 in Frontiers in Artificial Intelligence and Applications, 77-91. Amsterdam: IOS Press.

Fano, Roberto Mario. 1961. *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: MIT Press.

Fotzo, Hermine Njike, and Patrick Gallinari. 2004. "Learning «Generalization/specialization» Relations between Concepts: Application for Automatically Building Thematic Document Hierarchies." In *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, 143–155. Le Centre de Hautes Études Internationales D'informatique Documentaire.

Gherasim, Toader, Harzallah, Mounira, Berio, Giuseppe, and Pascale Kuntz. 2013. "Methods and Tools for Automatic Construction of Ontologies from Textual Resources: A Framework for Comparison and its Application." In *Advances in Knowledge Discovery and Management*, edited by Fabrice Guillet, Bruno Pinaud, and Gilles Venturini, 177–201. Berlin, Heidelberg: Springer.

Gómez-Pérez, Asunción, and David Manzano-Macho. 2004. "An Overview of Methods and Tools for Ontology Learning from Texts." *The Knowledge Engineering Review* 19(3): 187–212. doi: dx.doi.org/10.1017/S0269888905000251.

Gruber, Thomas. 1995. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing?" *International Journal of Human-Computer Studies* 43(5-6): 907–928.

Harris, Zellig. 1954. "Distributional Structure." *Word* 10(2–3): 146–162.

Hazman, Maryam, El-Beltagy, Samhaa, and Ahmed Rafea. 2011. "A Survey of Ontology Learning Approaches." *Database* 22(8): 36–43. doi:10.5120/2610-3642.

Hearst, Marti. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." Proceedings of the Fourteenth conference on Computational Linguistics, Vol. 2, 539–545. Association for Computational Linguistics. doi: dx.doi.org/10.3115/992133.992154.

IJntema, Wouter, Sangers, Jordy, Hogenboom, Frederik, and Flavius Frasincar. 2012. "A Lexico-semantic Pattern Language for Learning Ontology Instances from Text." *Web Semantics: Science, Services and Agents on the World Wide Web*, 15, 37–50.

Jurafsky, Daniel, and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, (2nd Ed.). Pearson/Prentice-Hall.

Lehmann, Jens, and Johanna Völker, eds. 2014. *Perspectives on Ontology Learning*. Amsterdam: IOS Press.

Maedche, Alexander, and Steffen Staab. 2001. "Ontology Learning for the Semantic Web." *IEEE Intelligent Systems* 16(2): 72–79. Doi: dx.doi.org/10.1109/5254.920602.

Meijer, Kevin, Frasincar, Flavius, and Frederik Hogenboom. 2014. "A Semantic Approach for Extracting Domain Taxonomies from Text." *Decision Support Systems* 62: 78–93. doi: dx.doi.org/10.1016/j.dss.2014.03.006.

Petasis, Georgios, Karkaletsis, Vangelis, Paliouras, Georgios, Krithara, Anastasia, and Elias Zavitsanos. 2011. "Ontology Population and Enrichment: State of the Art." In *Knowledge-driven Multimedia Information Extraction and Ontology Evolution*, edited by Georgios Paliouras, Constantine Spyropoulos, and George Tsatsaronis, 134–166. Berlin: Springer. doi: dx.doi.org/10.1007/978-3-642-20795-2_6.

Periñán-Pascual, Carlos. 2017. "Bridging the Gap within Text-data Analytics: A Computer Environment for Data Analysis in Linguistic Research." *Revista de Lenguas para Fines Específicos* 23(2): 111-132.

Periñán-Pascual, Carlos, and Francisco Arcas Túnez. 2010. "The Architecture of FunGramKB", 7th International Conference on Language Resources and Evaluation, Valletta (Malta). Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2667–2674.

Periñán-Pascual, Carlos, and Ricardo Mairal Usón. 2010. "La Gramática de COREL: Un Lenguaje de Representación Conceptual". *Onomázein* 21, 11–45.

Princeton University. 2010. "About WordNet." WordNet. Princeton University.

Shamsfard, Mehrnoush, and Ahmad Abdollahzadeh Barforoush. 2003. "The State of the Art in Ontology Learning: A Framework for Comparison." *The Knowledge Engineering Review* 18(4): 293–316.

Ureña Gómez-Moreno, Pedro, and Eva Mestre-Mestre. 2017. "Automatic Domain-specific Learning: Towards a Methodology for Ontology Enrichment." *Revista de Lenguas para Fines Específicos* 23(2):63–85. doi: dx.doi.org/10.20420/rlfe.2017.173.

Velardi, Paola, Faralli, Stefano, and Roberto Navigli. 2013. "Ontolearn Reloaded: A Graph-based Algorithm for Taxonomy Induction." *Computational Linguistics* 39(3): 665–707. doi: dx.doi.org/10.1162/COLI_a_00146.

Wong, Wilson, Liu, Wei, and Mohammed Bennamoun. 2012. "Ontology Learning from Text: A Look Back and into the Future." *ACM Computing Surveys* 44(4): 1–36. doi: 10.1145/2333112.2333115.

Zouaq, Amal, and Roger Nkambou. 2010. "A Survey of Domain Ontology Engineering: Methods and Tools." In *Advances in Intelligent Tutoring Systems*, edited by Roger Nkambou, Mizoguchi Riichiro, and Jacqueline Bourdeau, 103–119. Berlin, Heidelberg: Springer. doi: 10.1007/978364214363-2.

ANNEX I. Public health statement for *ethylbenzene* as published on the ATSDR website[5]

Note: the charts in this annex have been cropped for reasons of space and clarity.

What is ethylbenzene?

Colorless liquid that smells like gasoline	You can smell ethylbenzene in the air at 2 parts of ethylbenzene per million parts of air (2 ppm). It evaporates at room temperature and burns easily.
Used in industry and in consumer products	Ethylbenzene is found naturally in oil. Large amounts of ethylbenzene are produced in the United States; most of it is used to make styrene. Ethylbenzene is also used in fuels. Consumer products containing ethylbenzene include: • gasoline • paints and inks

What happens to ethylbenzene when it enters the environment?

Most commonly found in air	Ethylbenzene moves easily into the air from water and soil. Ethylbenzene in soil can also contaminate groundwater.
Rapidly broken down in air	*Air:* Ethylbenzene in air is broken down in less than 3 days with the aid of sunlight. *Water:* In surface water such as rivers and harbors, ethylbenzene breaks down by reacting with other compounds naturally present in water.

How might I be exposed to ethylbenzene?

Air	If you live in a city or near many factories or heavily traveled highways, you may be exposed to ethylbenzene in the air. Releases of ethylbenzene into the air occur from burning oil, gas, and coal and from industries using ethylbenzene. The median levels of ethylbenzene in air are: • 0.62 ppb in city and suburban locations

How can ethylbenzene enter and leave my body?

Rapidly enters your body	When you breathe air containing ethylbenzene, it enters your body rapidly and almost completely through your lungs. Ethylbenzene in food or water may also rapidly and almost completely enter your body through the digestive tract. It may enter through your skin when you come into contact with liquids containing ethylbenzene.

[5] https://www.atsdr.cdc.gov/PHS/PHS.asp?id=381&tid=66

How can ethylbenzene affect my health?

This section looks at studies concerning potential health effects in animal and human studies.

Short-term exposure in air	*Humans:* Exposure to high levels of ethylbenzene in the air for short periods can cause eye and throat irritation. Exposure to higher levels can result in vertigo and dizziness. *Animals*: Exposure to very high levels (about 2 million times the usual level in urban air) can cause death.
Long-term exposure in air	*Hearing:* Exposure to relatively low concentrations of ethylbenzene for several days to weeks resulted in potentially irreversible damage to the inner ear and hearing of animals.

How can ethylbenzene affect children?

This section discusses potential health effects in humans from exposures during the period from conception to maturity at 18 years of age.

Children are likely to have similar effects as adults	No information is available about the effects of exposure to ethylbenzene on children. It is likely that children would show the same health effects as adults. We do not know whether children will have effects at the same exposure levels as adults.
Birth defects	We do not know whether ethylbenzene causes birth defects in people. Minor birth defects and low birth weights have occurred in newborn animals whose mothers were exposed to air contaminated with ethylbenzene during pregnancy.

How can families reduce the risk of exposure to ethylbenzene?

Limit children's exposure to consumer products containing ethylbenzene	Use adequate ventilation to minimize exposure to ethylbenzene vapors from consumer products such as • gasoline • pesticides • varnishes and paints • newly installed carpeting • automobile products Sometimes older children sniff household chemicals in an attempt to get high. Your children may be exposed to ethylbenzene by inhaling products containing it, such as paints, varnishes, or gasoline. Talk with your children about the dangers of sniffing chemicals.
Store household chemicals out of reach of young children	Always store household chemicals in their original labeled containers out of reach of young children to prevent accidental poisonings. Never store household chemicals in containers children would find attractive to eat or drink from, such as old soda bottles.

Is there a medical test to determine whether I have been exposed to ethylbenzene?

Can be measured in blood and breath	Ethylbenzene can be measured in blood and in the breath of people exposed to ethylbenzene. This should be done within a few hours after exposure occurs because ethylbenzene's breakdown products (metabolites) leave the body very quickly.
Metabolites can be measured in urine	The presence of ethylbenzene's metabolites in urine might indicate that you were exposed to ethylbenzene; however, these breakdown products can also form when you are exposed to other substances, such as styrene.

What recommendations has the federal government made to protect human health?

The federal government develops regulations and recommendations to protect public health. Regulations can be enforced by law. The EPA, the Occupational Safety and Health Administration (OSHA), and the Food and Drug Administration (FDA) are some federal agencies that develop regulations for toxic substances. Recommendations provide valuable guidelines to protect public health, but cannot be enforced by law. The Agency for Toxic Substances and Disease Registry (ATSDR) and the National Institute for Occupational Safety and Health (NIOSH) are two federal organizations that develop recommendations for toxic substances.