# Building a Feature Taxonomy of the Terms Extracted from a Text Collection

Svitlana Moiseyenko, Alexander Vasileyko,
and Vadim Ermolayev

Department of Computer Science, Zaporizhzhia National University,
Zhukovskogo st. 66, Zaporizhzhia, Ukraine
svitlana.moiseyenko@gmail.com, vasileyko.alex@gmail.com,
vadim@ermolayev.com

**Abstract.** This position paper presents an approach for feature grouping and taxonomic relationship extraction with the further objective to build a feature taxonomy of a learned ontology. The approach needs to be developed as a part of the OntoElect methodology for domain ontologies refinement. The paper contributes a review of the related work in taxonomic relationships extraction from natural language texts. Within this review, the research gaps and remaining challenges are analyzed. The paper proceeds with outlining the envisioned solution. It presents the approach to this solution starting with the research questions, followed by the initial research hypotheses to be tested. Consequently, the plan of research is presented, including the potential research problems, the rationale to use and re-use existing components, and evaluation plan. Finally, the proposed solution, and the project, are placed in the broader context of the overall OntoElect workflow.

**Keywords:** feature taxonomy, relation extraction, subsumption, meronymy, ontology engineering, OntoElect.

## 1 Introduction

This paper presents an approach to building a taxonomy of the features, indicating the elicited requirements, within the OntoElect methodology [1]. OntoElect is an ontology refinement methodology, which consists of the following phases: (i) Feature Elicitation; (ii) Conceptualization and Formalization; (iii) Evaluation. Building the taxonomy of the developed or refined ontology is a major step within the Conceptualization and Formalization phase. It is composed of two steps: (i) Feature Grouping; and (ii) Relationship Extraction.

In this paper, we present our vision of an advanced approach for extracting specific types of relationships, namely subsumption and meronymy. These relationships are to be further used for building a feature taxonomy for the ontology under development or refinement. The approach is advanced as it combines the highlights of the existing approaches that are known from the related work. Due to this combination, we envision that the precision of the result will be higher than that of the State-of-the-Art.

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 59–70

The vision of the approach is presented as an M.Sci project proposal, as it: (i) fits as a part of the developed technology for the Feature Grouping phase of OntoElect; and (ii) may further be extended to extract other kinds of relationships to formalize the requirements for an ontology under development [1, 2].

The reminder of the paper is structured as follows. Section 2 reviews the related work on ontology learning from text and taxonomy extraction. Based on this review it outlines the research gaps. Sect. 3 presents our motivation to narrow the outlined research gaps by combining the advantages of the State-of-the-Art approaches with the knowledge we have already made available in the Requirements Elicitation phase of OntoElect. Research questions and initial research hypotheses are formulated. Sect. 4 presents our informed vision of the approach to developing the proposed solution and the plan of the project. Sect. 5 puts the proposed work and the envisioned solution in the broader context of the OntoElect project. Finally, we offer our conclusive remarks and a perspective view on the relevant future work in Section 6.

## 2 Related Work

Various taxonomy extraction approaches have been studied in the related work. These approaches, reviewed below, used different combinations of knowledge sources for extracting taxonomic relationships. Based on the combinations of these sources, the related work could be grouped into pattern-based, linguistic, statistical, graph-based, terminology-based, and dynamic clustering types of methods.

The pioneering works on pattern-based approaches to taxonomy extraction were [3, 4]. However, the patterns, proposed initially, were too specific to cover all the required true positive cases that can be met in natural language texts. Therefore, further research in the pattern-based direction looked into devising more flexible and generalizing patterns, following and extending [5, 6, 7]. Some authors explored the opposite direction and developed more specific “doubly-anchored” patterns [8].

Some results have been reported in combining linguistic and statistical approaches to taxonomic relationships extraction. It was reported that such a combination of the appropriate techniques allows increasing the precision of results [9, 10, 11]. Some other works exploited negative information within a linguistic approach [12]. A notable cluster of research followed a hybrid approach and combined NLP and pattern-based techniques, resulting in the proposal of lexico-syntactic patterns (LSP). LSP were used for linguistic and statistical matching in retrieving concepts and their relationships [13].

Some authors have noted that the use of plain texts for retrieving relationships does not yield required quality. As an additional source of quality information, they proposed to exploit an existing lexical resource, WordNet [22] or other Web resources for validating the relationships extracted from text [6, 14]. Further, graph-based techniques in NLP were used to parse textual knowledge sources and evaluate the obtained results using Word-Net [4].

A cluster of contributions adding to the quality of relationship extraction used additional knowledge provided by the terms extracted from texts, or text collections / corpora [15, 16, 6, 1, 17, 18]. Some of the works in this category used post-processing to improve recall and precision. This post-processing was done by applying: (i) statistics-based cuts [19, 20]; (ii) domain specificity / significance scores [1, 17, 21].

In addition to the reviewed typology of techniques, it needs to be mentioned that, broadly, the methods for taxonomy extraction could be classified as (i) unsupervised; (ii) semi-supervised; and (iii) supervised. Unsupervised methods are domain-neutral as these techniques do not require a domain-specific training set or other bits of expertize (e.g. specific rules) that help improve the quality of extraction. For instance, the feature grouping technique [18] is domain neutral as it is based on the partial matching of candidate term strings using a selection of string similarity measures. Supervised methods are mainly based on the use of domain-specific machine learning models. Topical representatives of this category are Word2Vec [23], GloVe [24], and SensEmbed [25]. Importantly, these supervised methods exploit terms embeddings in the text which are important source of knowledge about relationships and the contexts of terms. Semi-supervised methods were considered in [8] based on a handful of labeling and model learning techniques to extract categories (concepts) and relations from web pages. A semi-supervised algorithm was developed [4], that uses a root concept and recursive surface pattern to learn automatically from the Web relation pairs.

Finally, an early stage work using dynamic clustering for retrieving taxonomic relations (a supervised approach) was proposed in [26]. In its idea, this work resembles our background work on feature grouping [18]. The difference is that [18] groups terms and devises taxonomic relationships based on candidate strings matching.

Source-wise, all the reviewed approaches exploit only a part of the available information to improve the quality of taxonomy extraction. The proposal of this position paper is to develop a method that exploits all of the available types of knowledge: terminology with significance scores; LSP of term definitions; term embeddings in the text; available commonsense lexical databases (like WordNet).

## 3 Motivation, Background, Research Questions, and Hypotheses

The motive to develop the hybrid approach envisioned in this paper is the extraction of a feature taxonomy, in an unsupervised way, with higher quality than achieved currently by the State-of-the-Art techniques (Section 2). Our premise is that using all the available sources of information about potential taxonomic relationships will result in filtering out false positives while keeping the true positives not negatively affected. This will increase the balanced F-measure of our extraction – so the quality will be higher. A pictorial representation of the envisioned approach is given in Fig. 1.

Our approach exploits the background knowledge of the OntoElect project[1] in extracting a saturated [27] bag of terms from the document collection using the descending citation frequency ordering of the documents in the collection to balance terminol-

[1] https://www.researchgate.net/project/OntoElect-a-Methodology-for-Domain-Ontology-Refinement

ogy drift in time and the difference in the terminological impact of individual documents [28]. Term extraction provides a flat ordered list of the terms with their significance values. In OntoElect, these terms are further termed as features identifying the requirements for the refined ontology. This flat list of features could further be transformed in a set of hierarchical groupings [18] hinting about existing subsumption or meronymy relationships between some of the terms. These background-based steps are numbered as (1) and (2) in Fig. 1.

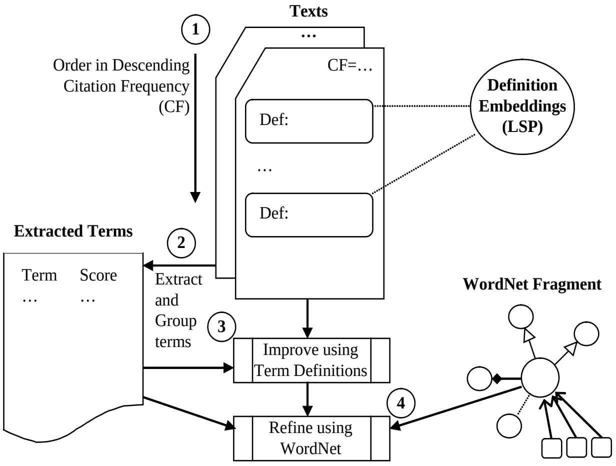

**Fig. 1**. The vision of the processing pipeline for taxonomy extraction and refinement

According to our background knowledge, a draft taxonomy built of the groupings (step 2) is of poor quality. The reason for that is that only the information about approximate string matches is used for grouping. Hence, the refinement is required. In this context, our first research question is: (**RQ1**) Are there more bits of relevant information in the text collection used for term extraction at step (2) that may help refine the feature taxonomy built using grouping?

Our corresponding research hypothesis (step (3) in Fig. 1) is:

(**H1**) The feature taxonomy may be improved by:

(i) Extracting term definition **embeddings** from the documents containing the involved **terms** and having the highest **citation frequency** (CF), following a domain-neutral **rule-based** approach based on LSP [13]

(ii) Doing **Part of Speech** (PoS) tagging of these embeddings

(iii) Extracting potential **taxonomic relationships** from the PoS taggings

One more research question relevant regarding the quality aspect is (**RQ2**): Are all the relationships, extracted at step (3), taxonomic relationships; has the semantics been properly extracted? We propose to answer RQ2 by validating the extracted relationships using the general-purpose (commonsense) knowledge provided by WordNet. This is reflected as step (4) in Fig. 1. Suppose that an extracted taxonomic relationship is supported by an explicit representation of it in WordNet, regarding a similar term or its superclass. Then, such a support could be regarded as a commonsense evidence of the correctness of this relationship. Otherwise, if a relationship is not supported, the possibility of it to be a false positive grows higher. A weak aspect in this approach is that WordNet is not supposed to contain and represent professional terminology, in a subject domain. Hence, there is a risk that a relationship of a valid feature does not find its support. As a remedy, the significance score of the respective term could be evaluated. The higher the confidence score, the higher is the probability that the feature, and the relationship, are true positives. In this context, our research hypothesis is:

(**H2**) The quality of the taxonomic relationships in a feature taxonomy could be validated, in a balanced way, by:

(i) Seeking for the support of this relationship by the evidence provided by WordNet

(ii) Balancing the lack of the coverage of professional terminology in WordNet by high significance scores of respective terms retrieved at step (2)

## 4 Approach to Solution and Research Plan

The outline of the proposed solution for improving the quality of taxonomic relationships extraction from texts has been given in Sect. 3. In this section, we discuss, in more detail, the research problems that have to be solved on the approach to the working solution. This discussion is structured using the chosen methodology and thus represents our proposed research plan.

Methodologically, the proposed project follows the pattern of the Scientific Method [32] and approaches its objectives iteratively. Every iteration starts with the formulation or refinement of a research hypothesis. Our initial hypotheses have been formulated in Sect. 3 as H1 and H2. Within the subsequent phase of an iteration, the instruments for testing the hypothesis are materialized, e.g. software is implemented, adopted, or adapted, and dataset(s) prepared. Finally, the hypothesis is tested in an evaluation experiment and the results are assessed regarding their proximity to the project objective.

In the remainder of this section, we outline our vision of the first iteration following the abovementioned pattern of “hypothesize – materialize – evaluate”.

### 4.1 Potential Research Problems in the Processing Pipeline

Until now, we presented the envisioned approach to feature taxonomy extraction without looking into the technical details. In this subsection, we point out the parts of the processing pipeline, which may be problematic to develop without elaborating these technical details and, possibly, refining or revising our research hypotheses H1 and H2.

**Term groupings and false positive relationships** (H1). Our prior work on terms grouping [18] revealed that a pair of term candidate strings might have high similarity (0.85-0.95 out of 1.00) but carry similar (partial positives – PP) or fully different (partial negatives - PN) semantics. PP pairs are exactly the cases in which taxonomic relationships may be sought. However, PNs are the false positive pairs. We admitted in [18] that there was no reliable way to filter out PNs as the similarity value intervals overlap substantially for PPs and PNs. This was perhaps one of the major reasons for term grouping to yield low quality in taxonomic relationships extraction. To improve this quality, the analysis of the composition of the candidate strings might be helpful. If for example, the PP term strings are “time interval” and “unbounded time interval” then the second string in the pair differs from the first by the added word token “unbounded”. Hence, the addition of a word token might indicate that a taxonomic relationship exists in this pair. Furthermore, the added word token is an adjective. Therefore, the recognition of PPs could be improved if the information about the parts of speech is exploited.

**Patterns for reliably locating term definitions in text** (H1). We devise the rationale for using term definitions for extracting the relationships (and properties) of these terms from the analysis of the deliberation patterns used in defining things by humans. Indeed, as known from Cognitive Science, humans define things:

- Either **top-down**, by: (i) relating the defined thing to the known thing that is more abstract or general; and (ii) specifying the additional properties for the defined thing. For example, “an unbounded time interval is a time interval, whose endings are not fixed”.
- Or **bottom-up**, by collecting the things that are less abstract and finding their common properties. For example, “an unbounded time interval, an open time interval, a convex time interval are all time intervals as these are not instant in time, i.e. have duration”.

Our premise, in the context of H1, is that the parts of text that define terms could be distinguished from the rest of the text based on their structure. Hence, the patterns of this structure for all the ways of defining things have to be developed and tested. This might be not that straightforwardly easy as outlined above. One possible reason is that, even for a fixed way of defining things, there might be different styles of formulations in a natural language. For example, for a top-down definition, the following two definitions are correct but follow different stylistic patterns: (i) “an unbounded time interval is a time interval, whose endings are not fixed”; (ii) “a time interval, whose endings are not fixed is an unbounded time interval”.

**A proper size of a term embedding in a definition text fragment** (H1). The definitions of terms could be given in one sentence, but could also span across several consecutive sentences. Hence, an open question in the context of locating a term definition embedding is how broad, in sentences, is has to be. Currently, we do not envision any formal way to specify the threshold. It is planned that the valid samples of term definitions need to be collected and analyzed. Furthermore, different thresholds need to be tested in the evaluation experiments on “gold standard” datasets (Sect. 4.3).

**Weighting WordNet support and term significance scores** (H2). Intuitively, it might be straightforward that if a candidate string, which is not a stop word, frequently appears in the text then it might be a term used in the text. This is reflected by a high significance score of such a string. Consequently, if there are indications of a taxonomic relationship between two strings with high significance scores then our confidence in the existence of this relationship is ought to be high. It is excellent if such a relationship gets explicit support in a lexical resource, like WordNet. However, it is not always the case as terminologies belonging to specialist domains may not be included in these lexical resources. Therefore, an open question in this context is what sort of evidence prevails – high text colocation confidence or no explicit support in WordNet. In the proposed approach, it is planned to find the proper balance between these two kinds of evidence by applying linear weighting. The weights will be empirically determined in the experiments on the “gold standard” datasets.

### 4.2 Materialization of the Required Components

To implement the project pipeline, several research and development tasks have to be planned regarding algorithmic and software implementation.

Term grouping software, to be used at step (2) is available as our background as a proof-of-concept implementation [18]. One of the shortcomings of this implementation is that the run times are too big for realistic full text document collections. The plan is to optimize this software by applying an efficient multiple string matching algorithm [33].

For step 3, it is planned to develop the lexico-syntactic patterns based on the analysis of the grammatically parsed text fragment definitions. For text parsing and PoS tagging, Stanford CoreNLP[2] will be used as the core API. It allows automatically elaborating the structure of sentences in terms of phrases, parts of speech, and syntactic dependencies. Hence, we expect to receive initial indicators for detecting the definitions of features in a domain-neutral way. Based on these taggings and the analysis of their types at a sentence level, the rules will be developed for extracting taxonomic relations from the feature definition embeddings.

For manipulating the elements in the WordNet database, it is planned to use the LxicaDB WordNet HTTP API [3], which allows querying WordNet via http. In addition, an algorithm for measuring the impact of the external support, balanced with the allocation confidence, needs to be developed for the instrumental pipeline to support evaluation experiments.

### 4.3 Evaluation Methodology and Plan

The goal of the evaluation phase is to test the hypotheses and find out how close the results are to the expected state of affairs. For evaluation at each iteration, in addition to the materialized solution (Sect 4.2), evaluation objectives, methodology, plan, data, and execution environment have to be specified. In this section, we outline the components, which will presumably remain unchanged in iterations, like methodology. We also specify the rest for the initial hypotheses H1 and H2 (Sect. 3).

**Evaluation Objectives**. The evaluation objectives are formulated in a way to check if the tested hypotheses hold true. As for H1, the objective (**EO1**) is to measure the quality of the extraction of taxonomic relationships using the steps 2, 3, and 4 of the processing pipeline (Fig. 1) and find out if the measured quality increases from step to step. To test H2 the objective is twofold. The first sub-objective (**EO2.1**) is finding out if there is support for the extracted relationships in WordNet and measure the proportion of the supported versus unsupported taxonomic relationships. This will presumably

[2] https://stanfordnlp.github.io/CoreNLP/

[3] http://www.lexicadb.com/lxserver/wn_http.html

allow verifying the utility of WordNet as a source of information where such support could be sought. Secondly, and if the support found in WordNet is sufficient, the sub-objective (**EO2.2**) is to find out the proportion of the unsupported true positives, having high significance scores. Achieving this sub-objective may allow finding proper weighting for reaching the balance mentioned in H2.

**Evaluation Methodology**. The outlined evaluation objectives provide the rationale for choosing a proper methodology for experimental evaluation. EO1 suggests that it would be rational to organize the experiments using the pattern of an ablation study[4] in which the quality measurements after steps 4, 3, and 2 are compared. Both EO1 and EO2 are based on quality measurement. One of the mainstream approaches for measuring the quality of extraction is the use of a measure based on a combination of recall and precision. One of the most often used measures of this sort in information retrieval is balanced F-measure [31]. The problem with it is that true/false positives and negatives have to be reliably identified and counted in order to get the trusted value. This is done using an appropriate "gold standard" dataset. In the context of the presented project, in a "gold standard" dataset all valid terms/features, taxonomic relationships among the features, and a feature taxonomy have to be manually extracted by a domain expert. The result of this manual extraction is regarded as the ground truth, which is used to measure the quality of automated extraction. Furthermore, as the goal of the proposed research is developing a domain-neutral (unsupervised) solution, we need several "gold standard" datasets for evaluation.

**Evaluation Data**. Based on the evaluation methodology, we are seeking for several collections of professional documents in different subject domains. Each of these collections has to be processed by domain experts to be used as a "gold standard" dataset. The experts have to manually extract terms, taxonomic relationships, and build the feature taxonomy. This work, if done from scratch, is too laborious for being feasibly undertaken in an individual (Master) project. Therefore, an exploratory search for available manually pre-processed datasets have to be undertaken at the initial phase of the project. Currently, the following collections could be pointed to as the ones for which the required pre-processing has been done at least in part:

- The **TIME paper collection** and **Syndicated Ontology of Time** (SOT) [29, 1]. SOT is available as our background result in the OntoElect project. It contains the feature taxonomy manually built [29] based on the terms extracted from the TIME document collection[5]. Hence, the taxonomy and taxonomic relationships are the available parts of the "gold standard" for TIME and could be used as one of the "gold standard" datasets in our experimental evaluation.
- Gold standard taxonomies of the **SemEval** initiative, taxonomy extraction evaluation task [30]. SemVal offers several gold standard taxonomies in different domains together with the sets of extracted terms. One shortcoming of these data is that the

[4] Ablation study is the experimental procedure for testing deep neural networks in which certain parts of the network, e.g. layers, are gradually removed to understand the influence of each individual part on the overall result.

[5] TIME collection in plain texts: http://dx.doi.org/10.17632/knb8fgyr8n.1#folder-d1e5f2b6-c51e-4572-b10d-0e2ebccead02

source document collections are not provided. Secondly, the significance scores of the extracted terms are not provided. Hence, these datasets could be used for cross-evaluation at the final iterations of our research workflow.

**Evaluation Plan**. The plan of performing evaluation experiments is straightforwardly inferred from Fig. 1. In each experiment for each individual dataset: (i) the processing step (2, 3, then 4) is executed; (ii) in each step the quality of taxonomy extraction is measured (balanced F-measure) by comparing the result of extraction to the gold standard result.

**Execution Environment**. All the planned computations are sufficiently lightweight to be run on a conventional laptop computer in affordable time.

## 5 The Work in a Broader Context

The proposed research is targeted to fill in the research gap in the Conceptualization Phase of the OntoElect approach for domain ontology refinement. The goal of OntolElect is to provide the methods and instrumental support for the refinement of a domain ontology in an arbitrary subject domain. The idea of the approach is that the requirements for an ontology are extracted, from a complete document collection describing the subject domain, as terms (features) in the Requirements Elicitation phase. In the Conceptualization phase, these requirements are formalized as ontological fragments. For that, different features, representing concepts, are grouped and the feature taxonomy is built. Finally, the features, representing properties are added to the nodes in the feature taxonomy. Hence, the formalized requirements are formed ass ontological contexts [1]. The workflow is pictured in Fig. 2. The part of this workflow that covers the scope of the proposed work is given in the inner rounded rectangle with solid border. As it is seen from the sequence of tasks, the solution that will be developed implements the instrumental support for the steps of Feature Grouping and Categorization and Building the Feature Taxonomy.

## 6 Conclusive Remarks and Outlook

In this position paper, we outlined our vision of a focused M.Sci proposal, which is topically aligned with our plans in the OntoElect project. We analyzed the related work in the field of extracting taxonomic relationships from a natural language text. Based on this analysis, we outlined the research gaps in the form of the open research questions and formulated the relevant research hypotheses that emerged from our vision of the potential solution.

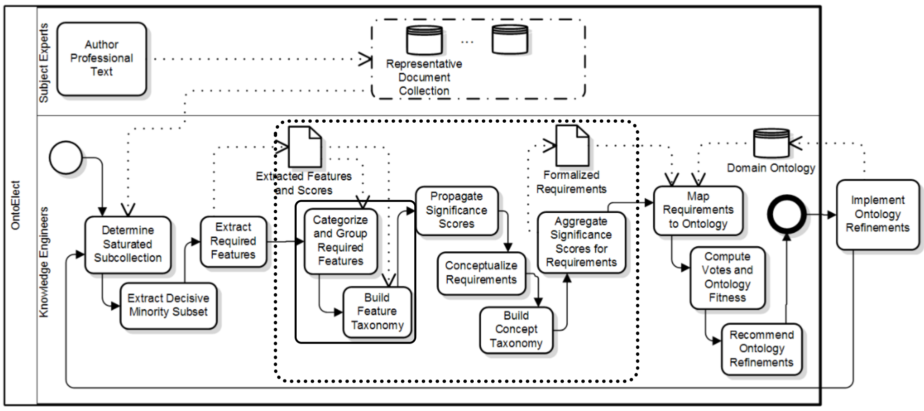
OntoElect
Subject Experts
Knowledge Engineers
Author Professional Text
Representative Document Collection
...
Extracted Features and Scores
Formalized Requirements
Domain Ontology
Determine Saturated Subcollection
Extract Decisive Minority Subset
Extract Required Features
Categorize and Group Required Features
Build Feature Taxonomy
Propagate Significance Scores
Conceptualize Requirements
Build Concept Taxonomy
Aggregate Significance Scores for Requirements
Map Requirements to Ontology
Compute Votes and Ontology Fitness
Recommend Ontology Refinements
Implement Ontology Refinements

**Fig. 2**: The Ontoelect workflow (adapted from [1]). Conceptualization phase is shown within the dotted rounded rectangle. The context of the proposed work is given in the inner rounded rectangle with solid border.

The idea of this proposal is to combine the use of more relevant sources of information about taxonomic relationships, than used in the State-of-the-Art solutions to date. Hence, the proposed approach is hybrid. We proposed to organize the envisioned solution as a four-step processing pipeline in a way to add more extraction quality at each consecutive step. We also presented our plan for performing this research that follows the pattern “hypothesize – materialize – evaluate” within each iteration. In the plan, we included our rationale for choosing the methodology, outlined potential research problems related to the steps of the proposed pipeline, and presented the way to perform experimental evaluation in the project.

Our planned future work related to the proposed project is extending the solution to the extraction of all types of relationships for incorporating properties into the ontological representations of requirements in the subsequent steps of the Conceptualization phase workflow.

## References

1. Ermolayev, V.: OntoElecting requirements for domain ontologies. The case of time domain. EMISA Int J of Conceptual Modeling **13**(Sp. Issue), 86–109 (2018). doi: 10.18417/emisa.si.hcm.9
2. Moiseyenko, S., Ermolayev, V.: Conceptualizing and formalizing requirements for ontology engineering. In: Antoniou, G., Zholtkevych, G. (eds.) PhD Symposium at ICTERI 2018, CEUR-WS, vol. 2122, pp. 35–44 (2018). online
3. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: 14th Conf on Computational Linguistics, pp. 539–545 (1992)
4. Kozareva, Z., Hov, E.: A semi-supervised method to learn and construct taxonomies using the web. In: Proc. 2010 Conf on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 1110–1118, MIT, Massachusetts, USA (2010)

5. Ritter, A., Soderland, S., Etzioni, O.: What is this, anyway: automatic hypernym discovery. In: AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read, pp. 88–93 (2009)
6. Tuan, L.A., Kim, J., Ng, S.K.: Taxonomy construction using syntactic contextual evidence. In: 2014 Conf on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 810–819 (2014)
7. Snow, R., Jurafsky, D., Ng. A.: Learning syntactic patterns for automatic hypernym discovery. In: 17th Annual Conf on Neural Information Processing Systems, pp. 1297–1304 (2005)
8. Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: 3d Int Conf on Web Search and Web Data Mining, pp. 101–110 (2010)
9. Diederich, J., Balke, W.-T.: The semantic growbag algorithm: automatically deriving categorization systems. In: Research and Advanced Technology for Digital Libraries, 11th European Conf, ECDL 2007, pp. 1–13 (2007)
10. Etzioni, O., Cafarella, M.J., Downey, D., Kok, S., Popescu A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates A.: Web-scale information extraction in knowitall (preliminary results). In: 13th Int Conf on World Wide Web, pp. 100–110 (2004)
11. Wu, W., Li, H., Wang, H.,Zhu, K.O.: Probase: a probabilistic taxonomy for text understanding. In: ACM SIGMOD Int Conf on Management of Data, pp. 481–492 (2012)
12. Wang, C., He, X.: Chinese hypernym-hyponym extraction from user generated categories. In: 26th Int Conf on Computational Linguistics, pp. 1350–1361 (2016)
13. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: a look back and into the future. ACM Comput. Surv. **44**(4), 20:1–20:36 (2012)
14. Tuan, L.A., Kim, J., Ng, S.K.: Incorporating trustiness and collective synonym/contrastive evidence into taxonomy construction. In: 2015 Conf on Empirical Methods in Natural Language Processing, EMNLP 2015, pp. 1013–1022 (2015)
15. Yang, H.: Constructing task-specific taxonomies for document collection browsing. In: 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1278–1289 (2012)
16. Zhu, X., Ming, Z., Zhu, X., Chua, T.: Topic hierarchy construction for the organization of multi-source user generated contents. In: 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval, pp. 233–242 (2013)
17. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.-E.: Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. In: Ermolayev, V., et al. (eds.) Revised Selected Papers of ICTERI 2013, CCIS, vol. 412, pp. 136–162 (2013). doi: 10.1007/978-3-319-03998-5_8
18. Kosa, V., Chaves-Fraga, D., Keberle, N., Birukou, A.: Similar terms grouping yields faster terminological saturation. In: Ermolayev, V. et al. (eds.) ICTERI 2018. Revised Selected Papers. CCIS, vol. 1007, pp. 43–70 (2019). doi: 10.1007/978-3-030-13929-2_3
19. Navigli, R.,Velardi, P.: Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics **30**(2), 151–179 (2004)
20. de Knijff, J., Frasincar, F., Hogenboom, F.: Domain taxonomy learning from text: the subsumption method versus hierarchical clustering. Data & Knowledge Engineering **83**, 54–69 (2013)
21. Alfarone, D., Davis, J.: Unsupervised learning of an IS-A taxonomy from a limited domain-specific corpus. In: 24th Int Joint Conf on Artificial Intelligence, pp. 1434–1441 (2015)
22. Fellbaum, C. (ed.): WordNet. An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)

23. Mikolov, T., Sutskever, I., Chen, K., Corrado G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: 27th Annual Conf on Neural Information Processing Systems, pp. 3111–3119 (2013)
24. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: 2014 Conf on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1532–1543 (2014)
25. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: learning sense embeddings for word and relational similarity. In: 53d Annual Meeting of the Association for Computational Linguistics and 7th Int Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 95–105 (2015)
26. Yamane, J., Takatani, T., Yamada, H., Miwa, M., Sasaki, Y.: Distributional hypernym generation by jointly learning clusters and projections. In: 26th Int Conf on Computational Linguistics, pp. 1871–1879 (2016)
27. Kosa, V., Chugunenko, A., Yuschenko, E., Badenes, C., Ermolayev, V., Birukou, A.: Semantic saturation in retrospective text document collections. In: Mallet, F., Zholtkevych, G. (eds.) ICTERI 2017 PhD Symposium, CEUR-WS, vol. 1851, pp. 1–8 (2017) online
28. Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Moiseyenko, S., Dobrovolskyi, H., Vasileyko, A., Badenes-Olmedo, C., Ermolayev, V., Corcho, O., and Birukou, A.: The influence of the order of adding documents to datasets on terminological saturation. Technical Report TS-RTDC-TR-2018-2-v2, Dept. of Computer Science, Zaporizhzhia National University, Ukraine (2018)
29. Ermolayev, V., Batsakis, S., Keberle, N., Tatarintseva, O., Antoniou, G.: Ontologies of time: review and trends. Int J of Computer Science and Applications **11**(3), 57–115 (2014)
30. Bordea, G., Buitelaar, P., Faralli, S., Navigli, R.: SemEval-2015 task 17: taxonomy extraction evaluation (TExEval). In: 9th Int. W-shop on Semantic Evaluation, SemEval 2015, pp. 902–910 (2015)
31. Manning, C. D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Cambridge University Press (2008)
32. Dodig-Crnkovic, G.: Scientific methods in Computer Science. In: Conf. for the Promotion of Research in IT at New Universities and at University Colleges in Sweden (2002)
33. Chowdhury, F. M., Farrell, R.: An efficient approach for super and nested term indexing and retrieval. arXiv preprint arXiv:1905.09761v1 [cs.DS], 23 May (2019)