Applying NLP in Text Extraction to develop Ontologies: Ontology Learning.

CSC5028Z - Mini Project

Willie Macharia - MCHWIL006
Department of Computer Science
University of Cape Town
South Africa
April 2021

Sharifa Negesa - NGSSHA002
Department of Computer Science
University of Cape Town
South Africa
April 2021

Abstract

Developing ontologies has been regarded as the most common way of representing knowledge within a certain domain area. To develop ontologies, various methodologies have been used by researchers over the years. These methodologies have been proposed and implemented to help ontologists develop ontologies that achieve high coverage of the concepts in the subject domain and relationships that exist between the domains. One of the methodologies that have been established is the use of text extraction to obtain both concepts and relationships that exist between the concepts. To extract text, techniques such as Natural Language Processing have been used over the years.

In this literature review, we review different approaches that NLP has used to extract text in the process of developing ontologies. The review focuses on the subtopic of ontology learning where various methods which have been used in different stages of ontology learning are highlighted. Trends and some existing ontology learning systems are discussed later in the review.

Keywords Ontology Learning, Natural Language Processing (NLP), Knowledge Base, Knowledge Acquistion, Concepts, Semantic Relation

1. INTRODUCTION

In the ontology development process, two main tasks are deemed important: discovering concepts in the domain area and discovering the relationships that exist between the concepts. Over the years, this process has been done manually which has turned out to be time-consuming, error-prone, and tiresome. This challenge has pushed researchers and ontologists to start developing techniques in which Natural Language Processing (NLP) can be applied to ontology development.

Natural language processing refers to the ability of a computational system to understand human language, that is, identifying key parts of language such as nouns, adjectives, verbs, and adverbs and making semantic meaning from the language. Human language can be in terms of voice, written text, or printed text. Ontology as a method of knowledge acquisition can rely on NLP to identify key concepts and relations between the concepts. NLP is a subfield of Artificial Intelligence where an AI model is well trained it can manage to extract keywords, key phrases, comprehend the meaning of the language, and in some cases generate a response [5]. The training of the model requires knowledge-rich documents to be used to improve the performance of the model during testing. To use NLP to create an ontology for a certain domain, knowledge-rich documents such as lexicons, standardized dictionaries, manuscripts, and thesauri are used to train NLP models to enable the models to understand the domain. The models are then used to generate various concepts that exist in that domain and the relationships between them.

Knowledge-rich sources can be in the form of text, audio, or images which all can be used in NLP [5]. However, In this literature review, we will limit our review to the use of text extraction by NLP systems to develop ontologies. Text extraction in ontology development can be used either in ontology learning or ontology population [21].

Ontology learning involves populating the TBox also called the Terminology box which consists of various classes and properties of a certain domain. Candidate terms are extracted from text using NLP and the term properties are also extracted by NLP.

Ontology population involves populating the ABox called the Assertion box which consists of various individuals that belong to various classes contained in the TBox. Various NLP systems that have been developed by researchers and use the ontology population to de-

velop ontology can be subdivided into three: rule-based systems, statistical approaches systems, and machine learning systems. The rule-based systems use lexico-syntactic patterns to determine the concepts and the relations between them [8]. They are developed manually by domain experts with collaboration from the ontologists where experts develop dictionaries about a certain domain and the ontologists use grammatical and syntactic rules to develop the ontology. Rules-based systems are not effective as they lack consistency due to a lack of depth analysis of the linguistics used [8]. Statistical systems use statistical methods such as fitness functions to detect the similarity of concepts and instances in an ontology. These systems aren't effective either because they aren't able to detect synonyms effectively. Machine learning systems deploy machine learning models to effectively identify concepts and instances in unstructured text. This has reduced human intervention in ontology development.

In this review, we have scaled down to focus on using the ontology learning method to develop ontologies. Ontology learning involves dividing the ontology development into various tasks namely: term extraction, synonym extraction, concept extraction, and relationship extraction. We will start by looking at the various methods used to extract terms in the text. We will then review how various researchers have managed to identify synonyms in various ontologies they have developed. We will then review how researchers have managed to filter concepts from terms. After identifying the concepts, we will then review how various relationships between concepts have been identified. After reviewing these four steps of ontology learning and how various researchers have done them, we will present a depth analysis on various metrics in the discussion section where we will compare various approaches and weigh which are better. Finally, we will conclude.

2. TERM EXTRACTION

Term extraction is a basic process in every ontology learning research whose aim is to get terms that may be considered as linguistic realizations of domain-specific concepts J. I. Toledo-Alvarado *et al* [30]. Term extraction comprises four text-mining approaches that include Linguistic, statistical, machine learning, and Hybrid. Linguamatics [2] describes text mining as an AI technology that uses natural language processing to transform free unstructured text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms. The various approaches of text mining are discussed in-depth as shown below;

2.1 Linguistics approach

Juan Antonio *et al* describes linguistic approaches as techniques that attempt to recover terms via linguistic pattern formation. A computer does Natural Language processing when it is doing an analysis based on the theory [1]. According to Terminology Coordination, [3] linguistic approach is an attempt to identify word combinations that match certain morphological or syntactic patterns (e.g. "adjective+noun" or "noun+noun"). It also stresses that linguistic term approaches work in a monolithic language and thus can not be used in a multilingual environment.

According to Raghu *et al* [6], the linguistic processing community has worked on techniques for understanding linguistic structures of sentences and applying these techniques for information extraction however, these approaches run into problems when the linguistic patterns conflict or lack completeness.

2.2 Statistical approach

Juan Antonio *et al* [22] describes statistical approaches as those that chiefly rely on external evidence presented through surrounding information. Statistical methods define whether a term is true and how it is related to a domain. Statistical methods are classified into two major categories that include the Machine learning method that attempts to treat the knowledge extraction as a classification process and the clustering approach that is based on a similarity measure [21]. There are two statistical measures i.e. unithood and termhood measures [12]. According to Thuy VU *et al* [31], unithood is the degree of strength or stability of syntagmatic combinations or collocations. Kageura *et al*[17] defines termhood as the degree that a linguistic unit is related to or represents domain-specific concepts. It is however noted that the distinction between these two terms that the term unithood focuses on only complex terms whereas termhood focuses on both simple and complex linguistic terms [17]. In the Automatic Term Recognition environment, term hood is termed as a measure of part or complete unit of frequency and in this case, an assumption is that complex terms can be made from existing simple terms.

According to Kageura *et al*[17] , the statistical findings that have been done by different researchers do not provide which is the best statistical measure of validity between the terms because their work depends explicitly or implicitly on assumptions concerning quantitative aspects of term usage. The limitations presented above make it unfavorable to consider the statistical approach as the best approach to term extraction because it does not directly apply to automated term recognition/term extraction but Information Retrieval as picking of adjacent words from a query is involved.

2.3 Machine Learning approach

With this approach, machine learning algorithms are used to extract instances from unstructured text [8]. In this approach, each word in the text is considered a candidate term whose classification predicts which words are terms of a specific domain [13]. It involves the text input preprocessing step where POS tagging of the corpora and normalization of the words of text. Normalization in this case helps to minimize the second Automated Term Extraction problems by allowing it to work with a lower Term Context representation dimensionality [13]. The machine learning approach has the ability to identify new candidate terms for populating and enriching ontologies but it can not do so without human intervention [8]. Conrado *et al* [13] presents merit for this approach since it facilitates the use of large Term Contexts and the ability that machine learning techniques learn by themselves to recognize a term is timesaving. From the above insights, a machine learning approach would only be a great approach if a researcher is not considering time as an important resource during automated term extraction else he/she would have to look for other alternatives to extract the terms.

2.4 Hybrid approach

According to Larranaga *et al* [12], Hybrid approaches combine both linguistic and statistical approaches of term extraction but rely on syntactic patterns for detecting candidate terms and use statistical measures to determine their domain-relatedness and relevance. Researchers have applied various techniques both supervised and unsupervised learning where unsupervised learning uses intelligent algorithms such as cognitive and linguistically driven analysis to extract concepts while the latter uses statistical and probability analysis, but none of the models has been perfect for those who work in a domain based on semantic analysis and thus a hybrid approach was adopted by many to generate good results and eliminate the limitations of both supervised and unsupervised learning techniques [19].

3. SYNONYM EXTRACTION

We define synonyms as words that have similar meanings. In rare cases, it is hard to find words that have exactly the same meaning in natural languages but there is rather a distinctive feature that exists among them. Synonym extraction has been done using the semantic dictionary or WordNet [4] over time, but there are not enough semantics till today because of the way humans communicate the algorithms in the semantic dictionary cannot find it. In this section we discuss how synonym extraction has been done using WordNet in details;

WordNet comprises three categories of semantic relationships such as Synonym, Hyponym, and Meronym. A hyponym is defined as a relation between words' meanings whereas a meronym is a part-of or part-whole or Has A relationship between words [26]. WordNet has been used as a measure of synonymy in many aspects because of its lexicon design and being a common designator of the natural languages [14]. Fellbaum [14] also elaborates that in the web structure of wordNet, words and synsets (group of words referring to the same concept) by conceptual-semantic and lexical relations.

Unlike adjectives, adverbs, verbs, nouns have been used by several researchers, philosophers, and linguists because the noun lexicon provides a fair way to look for gaps during comparisons of nouns in the WordNet and EuroNet. The article also elaborated that there would be generalization constraints as a result of lexicalization patterns because of the development of different multilingual WordNet, a case between the wordNet in French and the EuroNet. Fellbaum [14] believes that having a new perspective of WordNet could help solve such gaps.

Toledo-Alvarado *et al* [7] defines Synonym extraction as a process that facilitates term disambiguation where every term found on the corpus is assigned to a specific context for instance sun as a "Computer company" or the star thus classification of such terms in clusters using dictionaries and thesaurus helps in term disambiguation of those natural language meanings. However, it should be noted that similarity amongst words does not make them synonyms. Mohammed [24] addressed the synonymy challenge using the feed-forward neural network with backward propagation as a learning algorithm to capture semantic similarity of the words based on linear context. In the article, the findings showed that synonymy between adjectives occurred due to the functional similarity known as antonyms or domain similarity known as an association but not distributional similarity (words that occur in the neighborhood of a target). However, Mohammed [24] stresses that the challenge of using the neural technique is that similarity between vector word representations doesn't indicate synonymy and relatedness of the words.

When we contrast between using WordNet [4] and Neural techniques in Synonym extraction, we think WordNet is a better approach as it provides for scalability as well as helps to avoid the automatic disambiguation challenge which is a difficult problem in natural language processing [4].

4. CONCEPT EXTRACTION

Concept extraction can be defined as a process in the concept learning hierarchy algorithms in which text can be connected using potential relations that exist between them [7]. We looked at case studies in the clinical and legal concept extraction to determine how concept extraction has been done over time and provided our

thoughts on what we thought was a better approach for concept extraction.

The clinical concept extraction presents two approaches to which concepts can be extracted and these are rule-based and statistical approaches [15]. In our discussion, we concentrated mostly on the rule-based approach. The rule-based approach in this case deals with defining the rules that apply to selected patterns in a text and the application of morphological words to the text to obtain correct forms of concepts [29]. This approach is classified into rule-based, traditional machine learning, deep learning, or hybrid approaches [15]. This article elaborates that the implementation of Electronic Health Records (EHR) is dependent on the institutional infrastructure, system requirements, data usage agreements, research, and practice objectives.

We further delved to find out which of the rule-based approaches would be best during term extraction. In the article, it is noted that the rule-based approach allows for modification and refinement of the model based on existing implementation. This being the merit of this particular approach with less development and maintenance costs incurred, it presents demerits where development of the rules is substantial to manual effort and also the involvement of clinicians in the decision-making process is hard which makes its success questionable.

Another approach was the Machine learning approach which provides a shared task setting environment in which data that has been refined and evaluated is made available to the general public. It is also mentioned by Fu *et al* [15] that the Machine learning approach was used to patch deficiencies in entities with poor performance as extracted in the rule-based approach. The deep learning approach in this case was used to learn from the text in a sequential format in which it was presented. This approach is similar to the machine learning approach though, with deep learning, features such as word embeddings (e.g. Word2vec or GloVe) are difficult to use by the rule-based approaches [15].

The hybrid approach was a combination of rule-based and machine learning approaches into one system. This helped to improve on performance during the concept extraction process and also patch deficiencies extracted in the rule-based approach as earlier mentioned in the machine learning approach. In contrast, the rule-based approach was highly used compared to deep learning or machine learning. We however like to appreciate the fact that the hybrid approach of supervised machine learning would be the best method since it provides access to the task corpora to the public with high performance during the concept extraction process.

We also looked at the case study on the legal practices under the Statistical study of judicial practices journey [20]. According to this study, the syntex tool was used to identify terms and also establish syntactical differences among terms, and an example was seen in the French nationality whose objects were acquire and recover, the contexts of French nationality are thus (to acquire, OBJECT) and (to recover, OBJECT). Syntactic analysis in this case played an important role where it was easier to determine the semantic relations amongst terms from the syntactic relations. Furthermore,Guiraude Lame [20] presents the method of analysis of the coordination relations where terms are separated with "conjunctive phrases" and or "or". This is typically a tedious task since it relies on the fact that a program parses a document and determines if there is any conjunction that exists between them. It also needs manual intervention to determine which of the semantic relations do apply [20]. Comparing the two methods, we recommend that the method of syntactical analysis is much better as there is limited manual validation involved and it provides flexibility compared to the analysis of coordination of relations. It can also be noted that the analysis of coordination of relations is limited to only sections and subsections in a document and not their entire text which makes it a little ineffective to be used.

5. RELATIONSHIP EXTRACTION

Relationships that exist in an ontology can be classified into taxonomic relationships and non-taxonomic relationships. Taxonomic relationships are a type of relationship that exists between two concepts where one concept is "kind of" the other concept well known as "IS-A" relation in ontologies development [21]. Non-taxonomic relationships are the other relationships that exist between two concepts mainly defined as a binary predicate between two concepts. To extract these relationships, two main methods have been used so far by researchers: symbolic and statistical methods [21]. Symbolic methods are rule-based methods that use the lexico-syntactic rules or patterns to derive the semantic relationship between terms presented in the text while statistical methods use statistical analysis methods to provide the probability of existing relationships between concepts. To provide more analysis on what researchers have done on relationships extraction, we first focus on methods and procedures that have been used in extracting taxonomic relationships and then we focus on what has been done on extracting non-taxonomic relationships. In recent years, statistical methods have been extended to include machine learning methods such as clustering and deep learning neural network approaches[21].

5.1 Taxonomic Relationships

In-text, to obtain the taxonomic relationships, it is necessary to extract the hypernyms and hyponyms as they will form hierarchical relationships which translate to taxonomic relationships.

In 1992, Hearst MA [16], attempted to extract relationships between terms of interest by using lexico-syntactic patterns. In her research, Hearst gave a hypothesis that by identifying linguistic regularities present in a text corpus, syntactic relationships that exist between terms of interest can be extracted and used for ontology development or knowledge acquisition. To validate her hypothesis, she used a text corpus from Grolier's American Academic Encyclopedia where she searched for hyponym/hypernym relationships in the text. She managed to get 152 hypernym/hyponym relationships with corresponding 226 unique words. Hearst MA [16] used WordNet[4] to validate the results and found that out of 226 words, 180 words [16] had a hierarchy relationship in the WordNet. WordNet hierarchical relationships show that there existed taxonomic relationships between the terms which were extracted using the lexico-syntactic pattern. In conclusion, Hearst MA highlighted that this method could be used to extract relationships between text corpora that come from different domains and the method did not require a big knowledge base to extract the relationships. However, she highlighted that the method had a low recall performance meaning it is not a suitable method when an ontologist wants to develop a big and complex ontology. Despite concluding that the method could be used in text from different fields, Hearst MA did not test the method in a different field or different domain which renders her conclusion not correct.

Having identified that the lexico-syntactic method had a low recall, some researchers tried to improve the recall and applied the lexico-syntactic method to other domains. Caraballo [9] applied the noun coordination method to address the low recall challenge that Hearst had identified. Coordination in linguistics is a syntactic structure where two or more elements are conjoined also called conjuncts are linked [21]. Applying the coordination feature in extracting taxonomic relationships, Caraballo assumed that nouns in conjunction were semantically related and created a hierarchical tree of hypernyms. To extend Caraballo's work, Cedeberg and Widdows [10] used noun coordination to create a graph-based model that consists of noun-noun similarity which has been learned from noun coordination structure. Nouns were represented as the nodes of the graph while edges as the noun-noun relationships were represented. The edges were presented if two nouns existed in a coordination structure.

Recently, machine learning methods have been used to improve the performance of lexico-syntactic methods. Snow *et al* [28] used Hearst MA [16] patterns and created a dependency parse and searched for all the features that existed in each LSP path. The features obtained were then used to train a classifier which through testing, Snow was able to discover other new lexico-syntactic patterns.

5.2 Non-Taxonomic Relationships

To discover non-taxonomic relationships, Part-Of-Speech (POS) tagging has been used by various researchers who have taken the assumption that a concept connected with another concept via a verb or an adverb has a relation [21]. Kavalec *et al* [18], created a two-step process to extract non-taxonomic relationships. In the first step, he selected a verb denoted by the letter 'v' and two concepts denoted as c1 and c2 that occurred with a certain window of the verb. Second, he tabulated concept-concept-verb triples with their corresponding frequency. The triples with the highest frequency were then used to formulate a relationship between two given concepts. The frequency was a measure of co-occurrence of the pair concepts given a verb and sometimes two concepts may occur within a certain window of a verb and the two concepts have no mutual relationships at all. Therefore, this process did not provide a good methodology for extracting non-taxonomic relationships [18].

Chan *et al* [27] created a predicate recognition system to detect a description logic predicate that may be present in a natural language. The machine learning algorithm, Naive Bayes theorem was used to differentiate the previously learned predicates from the role of the attributes present in a class. An open-source tool called Link Grammar Parser was used to disassemble sentences and separate them into verb groups. The verb groups were then used to determine semantic links between the terms present in the sentences [27].

6. EXAMPLES OF ONTOLOGY LEARNING SYSTEMS

Several ontology learning systems have been developed in recent years to automate the ontology development from the text. These systems include OntoLearn, Text-To-Onto, Text-2-Onto, and Hasti. For this review we will discuss OntoLearn, Text-To-Onto and Text-2-Onto

OntoLearn [25] uses a hybrid approach of using statistical and linguistic methods to extract terms and complex terms from text corpora. The domain terms learned by OntoLearn are used to enrich WordNet [4] with the concepts extracted from the text. OntoLearn works on the principle of word sense disambiguation. The system has been evaluated in different domains

such as tourism, art and has achieved a recall rate ranging from 46% to 96% and a precision rate of over 65% [25].

Text-To-Onto [23] uses text mining techniques such as pruning techniques and association rules to extract terms, synonyms, concepts, relationships both taxonomic and non-taxonomic and instances present in text corpora. The unique element of Text-To-onto is the presence of many diverse algorithms which can be used for ontology pruning and refinement [23]. Evaluations of the system have achieved 76% accuracy of learning the non-taxonomic relationships. Few years after Text-To-Onto was developed, a new model was developed called Text-2-Onto [11] which was developed to address some of the challenges that Text-To-Onto faced. Key differences between Text-To-Onto and Text-2-Onto are: First, Text-2-Onto [11] represented the learned knowledge from the text as a Probabilistic Ontology Model(POM) which could be easily translated to other knowledge expressive languages such as OWL and RDF [11]. Text-2-Onto also incorporated dynamic changing POM which changed depending on the change of the text corpora and could allow the user to track the changes on the developed ontology concerning changes to the underlying text [11].

7. DISCUSSION

We have so far explored several methods that various researchers have used to carry out tasks in ontology learning. In this section, we compare and contrast the methods on the metrics of their size of the data, degree of human intervention, and customizability.

7.1 Size of the data

The rule-based methods such as the Hearst MA [16] lexico-syntatic patterns can only perform better where the text is small as using complex and huge data will not yeild good results as the pattern will be lost hence performing poorly. Statistical methods and machine learning approaches are suitable in scenarios where a lot of data is available. Text-2-Onto [11] is built on using machine learning approach and can handle a lot data and achieve a 76% accuracy.

7.2 Degree of human intervention

The degree of human intervention metric refers to the level of automation of different methods. The rule-based methods and statistical methods require a lot of human intervention to correct concepts and relations obtained from text compared to machine learning approaches. This is evident as ontology learning systems that have implemented machine learning approaches have achieved a high accuracy range compared to systems that have been developed using both rule based methods and statistical methods. For example DOODLEII [33] which uses statistical methods achieved a prescion rate of 30% while OntoLearn which uses machine learning methods achieved more than 50% precision rate.

It is noticeable that across all approaches be it statistical, rule-based or even machine learning, there is some level of human intervention needed []9. We think that however much an ontologist decides to use the Machine learning approach (if they have it in mind that everything will be computed for them), then they would be wrong because even the machine learning approach requires manually annotated examples to be able to execute a task as required. We would thus love to appreciate the role of machine learning engineers who facilitate the annotation process of the data because without such a hand, the methods are all prone to errors but a second eye through the entire process has been helpful.

7.3 Customizability

Customizability refers to the capacity in which a given model can be adapted when a particular concept has been changed. Here, we will look at three approaches that are shared across all the extraction techniques such as the rule-based approach, the machine-learning approach, and the hybrid approach. Unlike other methods, the rule-based approach allows a model to be modified and refined based on existing implementations. For instance in the Electronic Health Records(EHR) - based clinical concept extraction, several NLP systems such as MetMap, MedLEE, cTAKES, MedTagger, HiTex among others to identify clinical syndromes [32]. It is vividly seen that since the rule-based approach deals with only small data sets, customizability is easily achieved because any deficiencies that could be identified in the entities can easily be rolled out unlike with the machine learning approach which explicitly deals with large datasets making its extensibility tedious. We however believe a hybrid approach of both supervised machine learning and rule-based approach would be a somewhat better approach in extraction techniques because there will be a shared task setting that will prompt generalization of the annotated data where several researchers will be able to provide their thoughts and views on a specific domain.

8. CONCLUSIONS

Ontology learning has been regarded as an essential task in the development of ontologies over time. Considering the various text extraction techniques that have been described above, we believe that there is a lot of work that has to be done to close the gaps that arise while using different extraction techniques in text mining. The various literature that we have come through

has made us understand that none of the methods that we thought of as the best method was true because each method presents its constraint but is dependent on its purpose as may be described in the domain. Future work however should intend to consider identifying more terms specific to a domain and design more tools and methods to close the gaps that arise due to lexicalization, syntactic and semantic challenges.

9. TASKS ALLOCATION

The tasks allocation for writing literature review was done as team work where both Willie and Sharifa contributed equally to the writing, planning and researching about the topic. Progress meetings were done to ensure that deadline was met. Questions that rose during discussions were asked in class by both team members.

References

[1] Linguistic analysis explained - ascribe. `https://goascribe.com/linguistic-analysis-explained/`. (Accessed on 04/28/2021).

[2] What is text mining, text analytics and natural language processing? linguamatics. `https://www.linguamatics.com/`. (Accessed on 04/28/2021).

[3] Why terminology extraction? -. `https://termcoord.eu/2013/08/why-terminology-extraction/#:~:text=Term%20extraction%20tools%20using%20a,the%20content%20of%20the%20corpus`. (Accessed on 04/28/2021).

[4] Wordnet | a lexical database for english. `https://wordnet.princeton.edu/`. (Accessed on 04/28/2021).

[5] Al-Arfaj, A., and Al-Salman, A. Ontology construction from text: challenges and trends. *International Journal of Artificial Intelligence and Expert Systems 6*, 2 (2015), 15–26.

[6] Anantharangachar, R., Ramani, S., and Rajagopalan, S. Ontology guided information extraction from unstructured text. *arXiv preprint arXiv:1302.1335* (2013).

[7] Anoop, V., and Asharaf, S. Extracting conceptual relationships and inducing concept lattices from unstructured text. *Journal of Intelligent Systems 28*, 4 (2019), 669–681.

[8] Ayadi, A., Samet, A., de Beuvron, F. d. B., and Zanni-Merk, C. Ontology population with deep learning-based nlp: a case study on the biomolecular network ontology. *Procedia Computer Science 159* (2019), 572–581.

[9] Caraballo, S. A. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics* (1999), pp. 120–126.

[10] Cederberg, S., and Widdows, D. Using lsa and noun coordination information to improve the recall and precision of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* (2003), pp. 111–118.

[11] Cimiano, P., and Völker, J. A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)* (2005), Springer, pp. 227–238.

[12] Conde, A., Larrañaga, M., Arruarte, A., Elorriaga, J. A., and Roth, D. litewi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks. *Journal of the Association for Information Science and Technology 67*, 2 (2016), 380–399.

[13] Conrado, M., Pardo, T., and Rezende, S. O. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT student research workshop* (2013), pp. 16–23.

[14] Fellbaum, C. A semantic network of english: the mother of all wordnets. In *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, 1998, pp. 137–148.

[15] Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., et al. Clinical concept extraction: a methodology review. *Journal of Biomedical Informatics* (2020), 103526.

[16] Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics* (1992).

[17] Kageura, K., and Umino, B. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3*, 2 (1996), 259–289.

[18] Kavalec, M., and Svaték, V. A study on automated relation labelling in ontology learning. *Ontology Learning from Text: Methods, evaluation and applications*, 123 (2005), 44–58.

[19] Kumar, N., Kumar, M., and Singh, M. Automated ontology generation from a plain text using statistical and nlp techniques. *International Journal of System Assurance Engineering and Management 7*, 1 (2016), 282–293.

[20] Lame, G. Using nlp techniques to identify legal ontology components: concepts and relations. In *Law and the Semantic Web*. Springer, 2005, pp. 169–184.

[21] Liu, K., Hogan, W. R., and Crowley, R. S. Natural language processing methods and systems for biomedical ontology learning. *Journal of biomedical informatics 44*, 1 (2011), 163–179.

[22] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal 19*, 1-2 (2016), 59–99.

[23] Maedche, A., and Staab, S. Ontology learning for the semantic web. *IEEE Intelligent systems 16*, 2 (2001), 72–79.

[24] Mohammed, N. Extracting word synonyms from text using neural approaches. *Int. Arab J. Inf. Technol. 17*, 1 (2020), 45–51.

[25] Navigli, R., and Velardi, P. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics 30*, 2 (2004), 151–179.

[26] Sabrina, T., Rosni, A., and Enyakong, T. Extending ontology tree using nlp technique. In *Proceedings of National Conference on Research & Development in Computer Science REDECS* (2001), vol. 2001, Citeseer.

[27] Shu, C., Dosyn, D., Lytvyn, V., Vysotska, V., Sachenko, A., and Jun, S. Building of the predicate recognition system for the nlp ontology learning module. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (2019), vol. 2, IEEE, pp. 802–808.

[28] Snow, R., Jurafsky, D., and Ng, A. Y. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17* (2004).

[29] Szwed, P. Concepts extraction from unstructured polish texts: A rule based approach. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)* (2015), IEEE, pp. 355–364.

[30] Toledo-Alvarado, J., Guzman-Arenas, A., and Martínez-Luna, G. Automatic building of an ontology from a corpus of text documents using data mining tools. *Journal of applied research and technology 10*, 3 (2012), 398–404.

[31] Vu, T., Aw, A., and Zhang, M. Term extraction through unithood and termhood unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II* (2008).

[32] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics 77* (2018), 34–49.

[33] Yamaguchi, T. Acquiring conceptual relationships from domain-specific texts. In *Workshop on Ontology Learning* (2001).