PrOntoLearn: Unsupervised Lexico-Semantic Ontology Generation using Probabilistic Methods

Saminda Abeyruwan[1], Ubbo Visser[1], Vance Lemmon[2], and Stephan Schürer[3]

[1] Department of Computer Science, University of Miami, Florida, USA
{saminda,visser}@cs.miami.edu
[2] The Miami Project to Cure Paralysis, University of Miami Miller School of Medicine, Florida, USA
vlemmon@miami.edu
[3] Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Florida, USA
sschuerer@med.miami.edu

Abstract. Formalizing an ontology for a domain manually is well-known as a tedious and cumbersome process. It is constrained by the knowledge acquisition bottleneck. Therefore, researchers developed algorithms and systems that can help to automatize the process. Among them are systems that include text corpora for the acquisition. Our idea is also based on vast amount of text corpora. Here, we provide a novel unsupervised bottom-up ontology generation method. It is based on lexico-semantic structures and Bayesian reasoning to expedite the ontology generation process. We provide a quantitative and two qualitative results illustrating our approach using a high throughput screening assay corpus and two custom text corpora. This process could also provide evidence for domain experts to build ontologies based on top-down approaches.

Keywords: Ontology Modeling, Ontology Learning, Probabilistic Methods

1 Introduction

An ontology is a formal, explicit specification of a shared conceptualization [10], [22]. Formalizing an ontology for a given domain with the supervision of domain experts is a tedious and cumbersome process. The identification of the structures and the characteristics of the domain knowledge through an ontology is a demanding task. This problem is known as the knowledge acquisition bottleneck (KAB) and a suitable solution presently does not exist.

There exists a large number of text corpora available from different domains (e.g., the BioAssay high throughput screening assays[4]) that need to be classified into ontologies to faciliate the discovery of new knowledge. A domain of discourse

[4] http://bioassayontology.org/

(i.e., sequential number of sentences) shows characteristics such as 1) redundancy 2) structured and unstructured text 3) noisy and uncertain data that provide a degree of belief 4) lexical disambiguity, and 5) semantic heterogeneity problems. We discuss in depth the importance of these characteristics in section 3. Our goal in this research is to provide a novel method to construct an ontology from the evidence collected from the corpus. In order to achieve our goal, we use the lexico-semantic features of the lexicon and probabilistic reasoning to handle the uncertainty of features. Since our method is applied to build an ontology for a corpus without domain experts, this method can be seen as an unsupervised learning technique. Since the method starts from the evidence present in the corpus, it is can be seen as a reverse engineering technique. We use WordNet[5] to handle lexico-semantic structures, and the Bayesian reasoning to handle degree of belief of an uncertain event. We implement a Java based application to serialize the learned conceptualization to OWL DL[6] format.

The rest of the paper is organized as follows: section 2 provides a broad investigation of the related work. Section 3 provides details of our research approach. Section 4 provides a detail description of the experiments based on three different text corpora and the discussion. Finally, section 5 provides the summary and the future work.

2 Related Work

The problem of learning a conceptualization from a corpus has been studied in many disciplines such as machine learning, text mining, information retrieval, natural language processing, and Semantic Web. Table 1 shows the pros and cons of different techniques to solve the problem of *ontology learning*. Each method covers some portion of the problem and each method learns the conceptualization from terms, and present it as taxonomies and axioms to an ontology. On the other hand, most of the methods use a top-down approach, i.e., an initial classification of an ontology is given. The uncertainty inherited from the domain is usually dealt with by a domain expert, and the conceptualization is normally defined using predefined rules or templates. These methods show the characteristics of a semi-supervised and a semi-automated learning paradigm.

3 Approach

Our research focuses on an unsupervised method to quantify the degree of belief that a grouping of words in the corpus will provide a substantial conceptualization of the domain of interest. The degree of belief in world states influences the uncertainty of the conceptualization. The uncertainty arises from partial observability, non-determinism, laziness and theoretical and practical ignorance [19]. The partial observability arises from the size of the corpus. Even though

[5] http://wordnet.princeton.edu/
[6] http://www.w3.org/TR/owl-guide/

Table 1. The summary of the related work. Probabilistic learning (PR), never ending language learning (NELL), discovery and aggregation of relations in text (DART), recognizing textual entailment (RTE), automated theorem proving (ATP), natural language understanding (NLU), formal concept analysis (FCA), and ontology population (OP).

Work	Purpose	T-Box	A-Box	Method
PR [9], [12], [14] and [17]	reasoning	available	available	prob. theory
NELL [3]	24×7 learning	fixed	dynamic	ML techniques
DART [7]	world knowledge	$\times$	$\times$	semi-automated
RTE [2], and [13]	entailment	$\times$	$\times$	ATP
NLU [20]	commonsense rules	$\times$	$\times$	semi-supervised
Text2Onto [6]	ontology learning	$\surd$	$\surd$	semi-supervised
LexO [24]	complex classes	$\surd$	$\times$	semi-supervised
FCA [5]	taxonomy	$\surd$	$\times$	FCA
OP [4], and [23]	ontology population	available	available	semi-/supervised

a corpus many be large, it might not contain all the necessary evidence of an event of interest. A corpus contains ambiguous statements about an event that leads to a non-determinism of the state of the event. The laziness arises from the too much work that needs to be done in order to learn exceptionless rules and it is too hard to learn such rules. The theoretical and practical ignorance arises from lack of complete evidence and it is not possible to conduct all the necessary tests to learn a particular event. Hence, the domain knowledge, and in our case the domain conceptualization, can at best provide only a degree of belief of the relevant groups of words. We use probability theory to deal with the degrees of belief. As mentioned in [19], the probability theory has the same ontological commitment as the formal logic, though the epistemological commitment differs. The process of learning and presenting a probabilistic conceptualization is divided into four phases as shown in Figure 1. They are, 1) pre-processing 2) syntactic analysis 3) semantic analysis, and 4) representation.

3.1 Pre-processing

A corpus contains a plethora of structured and unstructured sentences. A lexicon of a language is its vocabulary built from lexemes [11], [15]. A lexicon contains words belonging to a language and in our work individual words from the corpus. In pure form, the lexicon may contain words that appear frequently in the corpus but have little value in formalizing a meaningful criterion. These words are called stop words or in our terminology: negated lexicon, and they are excluded from the vocabulary. We, first, part-of-speech tagged the corpus with the Penn Treebank English POS tag set [16]. We use the subset of tagset NN, NNP, NNS, NNPS, JJ, JJR, JJS, VB, VBD, VBG, VBN, VBP, and VBZ. The word length W_L above some threshold W_{L_T} is also considered. The length of a word, with respect to

POS context, is the sequence of characters or symbols that made up the word. By default, we consider that a word with $W_L > 2$ sufficiently formalizes to some criterion.

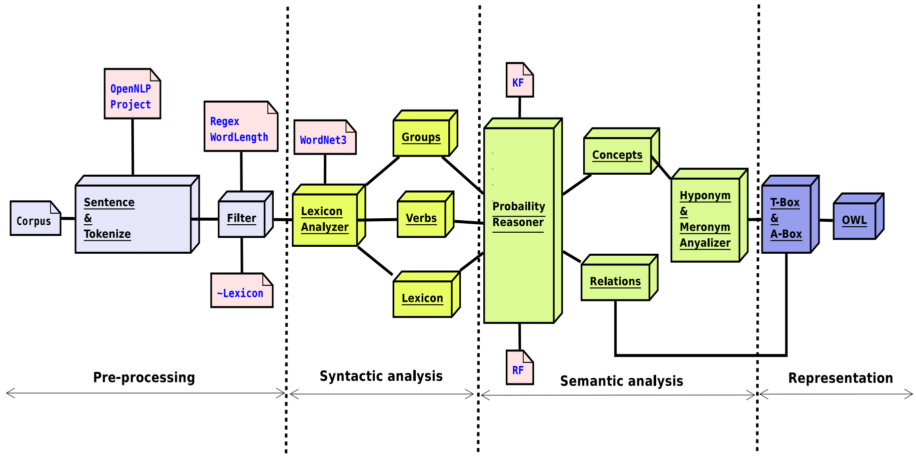

Fig. 1. Overall process: process categorizes into four phases; pre-processing, syntactic analysis, semantic analysis & representation

The pure form of the lexicon might contain words that need to be further purified according to some criterion. We use regular expressions for this task. Then we normalize and case-fold the words [15]. In addition to this there are families of derivationally related words with similar meanings. We use stemming and lemmatization to reduce the inflectional forms and derivational forms of a word to a common base form [15]. We achieve this with the aid of WordNets' stemming algorithms. We couple the knowledge of POS tag of the word to get the correct context when finding the common base form.

3.2 Syntactic Analysis

The primary focus on this phase is to look at the structure of the sentences and learn the associations among the vocabulary. We assume that each sentence of the corpus follows the POS pattern 1. 1,

$$(Subject_{NounPhrase}+)(Verb+)(Object_{NounPhrase}+) \tag{1}$$

We hypothesize that the associations learned from this phase provides the potential candidates for concepts and relations of the ontology. But the vocabulary itself does not provide sufficient ontology concepts. We use a notion of grouping of consecutive sequence of words to form an OWL concept. This grouping is done using an appropriate N-gram model [1]. We illustrate this idea using Figure 2.

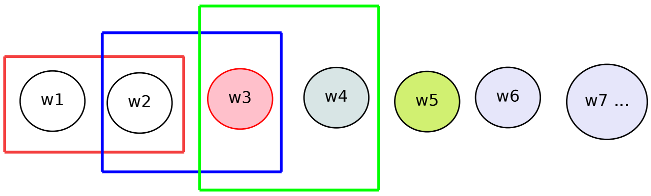

Fig. 2. An example three-gram model

The group $w_1 \circ w_2$ forms a potential concept in the conceptualization. We use the notation $x \circ y$ to show that the word y is appended to the word x. The groups $w_2 \circ w_3$, $w_3 \circ w_4$ etc. form other potential concepts in the conceptualization. Word w_3 comes after group $w_1 \circ w_2$. According to the Bayes viewpoint, we collect information to estimate the probability $P(w_3|\{w_1 \circ w_2\})$, which will be used to form IS-A relationships, $w_1 \circ w_2 \sqsubseteq w_3$ using an independent Bayesian network with conditional probability $P(\{w_1 \circ w_2\}|w_3)$. In addition to this, we count the groups appear in the left hand side and the right hand side of the expression 1 and the association of of these groups given the verbs. These counts are used in the third phase to create the relations among concepts.

3.3 Semantic Analysis

This phase conducts the semantic analysis with probabilistic reasoning, which constitutes the most important operation of our work. This phase determines the conceptualization of the domain using a probability distribution for IS-A relations and relations among the concepts. Our main definition of concept learning is given in Definition 1.

Definition 1. *The set $W = \{w_1, \ldots, w_n\}$ represents words of the vocabulary and each w_i has a prior probability $\theta_i > \tau$. τ is a prior threshold, which is known as the knowlege factor. The set $G = \{g_1, \ldots, g_m\}$ represents N-gram groups learned from the corpus and each g_j has a prior probability η_j. When $w \in W$ and $g \in G$, $P(w|g)$ is the likelihood probability π learned from the corpus. The entities w and g represent the potential concepts of the conceptualization and the set W provide the potential super-concepts of the conceptualization. Within this environment, an IS-A relationship between w and g is given by the posterior probability $P(g|w)$ and this is represented with a Bayesian network having two nodes w and g and is modeled by the equation,*

$$P(g|w) = \frac{\pi \times \eta}{\sum_i p(w|g_i) \times p(g_i)}. \tag{2}$$

Using the Definition 1, the probabilistic conceptualization of a domain is defined as follows.

Definition 2. *The probabilistic conceptualization of the domain is represented by an n-number of independent Bayesian networks sharing groups.*

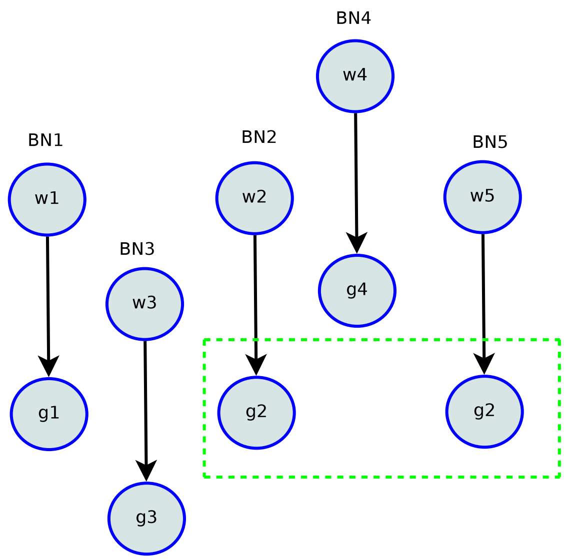

Fig. 3. w_1, w_2, w_3, w_4 and w_5 are super-concepts. g_1, g_2, g_3 and g_4 are candidate sub-concepts. There are 5 independent Bayesian networks. Bayesian networks 2 and 5 share the group g_2 when representing the concepts of the conceptualization

Figure 3 shows a simple example of the Definition 2. The interpretation of Definition 2 is: Let a set G contains an n-number of finite random variables $\{g_1, \ldots, g_n\}$. There exist a group g_i, which is shared by m words $\{w_1, \ldots, w_m\}$. Then, with respect to the Bayesian framework, BN_i of $P(g_i|w_i)$ is calculated and $max(P(g_i|m_i))$ is selected for the construction of the ontology. This means that if there exists two Bayesian networks and the Bayesian network one is given by the pair w_1, g_1 and the Bayesian network two is given by the pair $\{w_2, g_1\}$ then the Bayesian network that has the most substantial IS-A relationship is obtained through $max_{BN_i}(P(g_1|w_1))$ and this network is retained and other Bayesian networks will be ignored when building the ontology. If all $P(g_1|w_1)$ remains equal, then the Bayesian network with the highest super-concept probability will be retained. These two conditions will resolve any naming issues.

The next step is to induce the relationships to complete the conceptualization. In order to do this, we need to find semantics associated with each *verb*. We hypothesize that relations are generated by the verbs and the definition is as follows.

Definition 3. *The relationships of the conceptualization are learned from the syntactic structure model by the expression 1 and the semantic structure model by the lambda expression $\lambda obj.\lambda sub.Verb(sub, obj)$, where β-reduction is applied for obj and sub of the expression 1. If there exists a verb V between two groups of concepts C_1 and C_2, the relationship of the triple (V, C_1, C_2) is written as $V(C_1, C_2)$ and model with conditional probability $P(C_1, C_2|V)$. The Bayesian* *network for relationship is and the model semantic relationship is given by,*

$$P(C_1, C_2|V) = p(C_1|V)p(C_2|V) \rightarrow V(C_1, C_2)$$

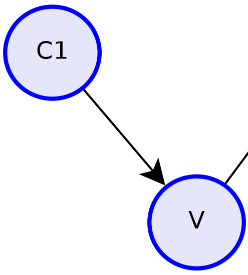

Fig. 4. Bayesian networks for relations modeling. C_1 and C_2 are groups and V is a verb

The relations learned from Defintions 3 needs to be subjected to a lower bound. The lower bound is known as the *relations factor*. When the corpus is substantially large, the number of relations is proportional to the number of verbs. Not all relations may relevant and the factor is used as the threashold. A verb may have antonyms. If a verb is associated with some concepts and these concepts happen to be associated with a antonym, the verb with the highest Bayesian probability value is selected for the relations map and the other relationship will be removed. Finally, the probabilistic conceptualization is serialized as an OWL DL ontology in the representation phase.

Our implementation of the above phases is based on Java 6 and it is named as `PrOntoLearn` (Probabilistic Ontology Learning).

4 Experiments

We have conducted experiments on three main data corpora, 1) the PCAssay, of the BioAssay Ontology (BAO) project, Department of Molecular and Cellular Pharmacology University of Miami, School of Medicine 2) a sample collection of 38 PDF files from ISWC 2009 proceedings, and 3) a substantial portion of the web pages extracted from the University of Miami, Department of Computer Science[7] domain . We have constructed ontologies for all three corpora with different parameter settings.

The first corpus contains high throughput screening assays performed on various screening centers. This corpus grows rapidly each month. We specifically limited our dataset to assays available on the 1^{st} of January 2010. Table 2 provides the statistics of the corpus. We extract the vocabulary generated

[7] http://www.cs.miami.edu

from *[a-zA-Z]+[-_]?\w** regular expression, and normalized them to create the vocabulary.

Table 2. The PCAssay (the BioAssay Ontology project) corpus statistics

Title	Statistics	Description
Documents	1,759	All documents are XHTML formated with a given template
Unique $ConceptWords$	13,017	Normalized candidate concept words from NN, NNP, NNS, JJ, JJR & JJS using [a-zA-Z]+[-_]?\w*
Unique $Verbs$	1,337	Normalized verbs from VB, VBD, VBG, VBN, VBP & VBZ using [a-zA-Z]+[-_]?\w*
Total $ConceptWords$	631,623	
Total $Verbs$	109,421	
Total Lexicon	741,044	$Lexicon = ConceptWords \bigcap Verbs$
Total $Groups$	631,623	

The average file size of the corpus is approximately 6 Kb. We conducted these experiments in a Genuine Intel(R) CPU 585 @ 2.16GHz, 32 bits, 2 Gb Toshiba laptop. It is found that the time required to build the conceptualization grows linearly. We use precision, recall and F1 measures to evaluate the ontology and recommendations from domain experts, specially to get comments on the generated bioassay ontology. The ontology that is generated is too large to show in here.Instead, we provide a few distinct snapshots of the ontology with the help of Protégé OWLViz plugin. Figures 5 and 6 show snapshots of the ontology created from the BioAssay Ontology corpus for input parameters $KF = 0.5$, N-gram = 3, and $RF = 0.9$. Figure 5 shows the IS-A relationships and Figure 6 shows the binary relationships.

According to experts, the ontology contains rich set of vocabulary, which is very useful for top-down ontology construction. The experts also mentioned that the ontology has good enough structure. The *www.cs.miami.edu* corpus is used to calculate quantitative measurements. The gold standard based approaches such as precision (P) and recall (R) and F-measure (F_1) are used to evaluate ontologies [8]. We use a slightly modified version of [21] as our reference ontology. Table 3 shows the results. The average precision of the constructed ontology is approximately 42%. It is to be noted that we use only one reference ontology. If we use another reference ontology the precision values varies. This means that the precision value depends on the available ground truth.

The results show that our method creates an ontology for any given domain with acceptable results. This is shown in the precision value, if the ground truth

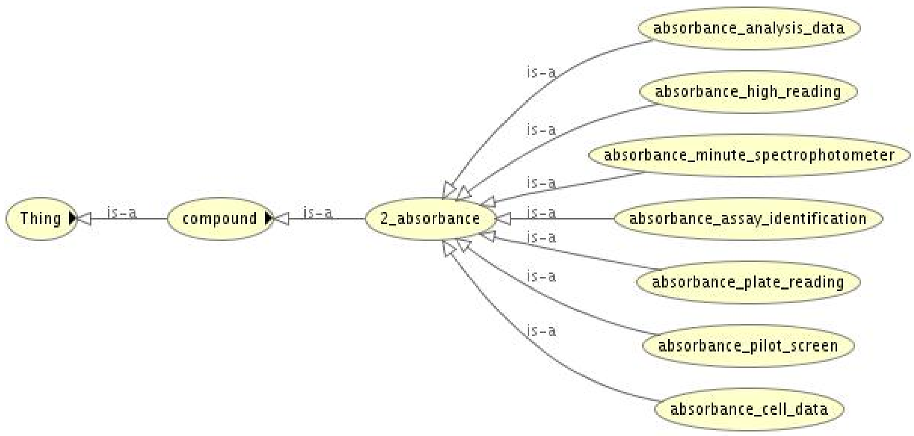

Fig. 5. An example snapshot of the BioAssay Ontology corpus with IS-A relations

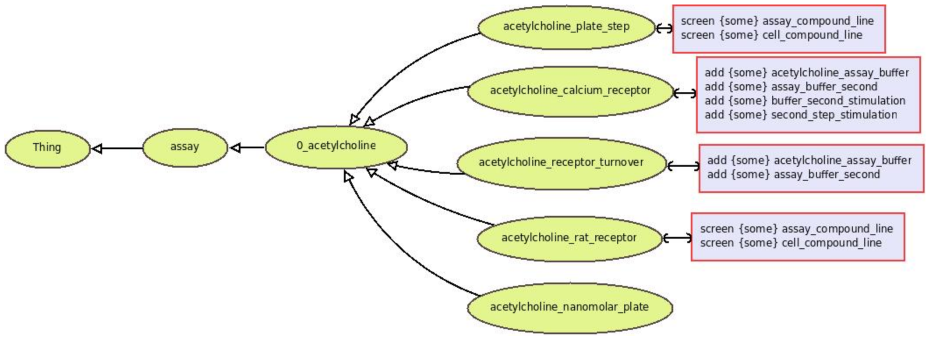

Fig. 6. An example snapshot of the BioAssay Ontology corpus with binary relations

Table 3. Precision, recall and F1 measurement for N-gram=4 and RF=1 using extended reference ontology

KF	Precision	Recall	F1
0.1	0.424	1	0.596
0.2	0.388	1	0.559
0.3	0.445	1	0.616
0.4	0.438	1	0.609
0.5	0.438	1	0.609
0.6	0.424	1	0.595
0.7	0.415	1	0.587
0.8	0.412	1	0.583
0.9	0.405	1	0.576
1.0	0.309	1	0.472

is available. On the other hand, if the domain does not have ground truth the results are subject to domain expert evaluation of the ontology. One of the potential problems we have seen in our approach is search space. Since our method is unsupervised, it tends to search the entire space for results, which is computationally costly. We thus need a better method to prune the search space so that out method provide better results. According to domain experts, our method extracts good vocabulary but provides a flat structure. They have proposed a sort of a semi-supervised approach to correct this problem, by combining the knowledge from domain experts and results produced by our system. We left the detailed investigation for future work.

Since our method is based on the Bayesian reasoning (which uses N-gram probabilities), it is paramount that the corpus contains enough evidence of the redundant information. This condition requires that the corpus to be large enough so that we can hypothesize that the corpus provides enough evidence to build the ontology.

We hypothesize that a sentence of the corpus would generally be subjected to the grammar rule given in expression 1. This constituent is the main factor that uses to build the relationships among concepts. In NLP, there are many other finer grained grammar rules that specifically fit for given sentences. If these grammar rules are used, we believe we can build a better relationship model. We have left this for future work.

At the moment our system does not distinguish between concepts and the individuals of the concepts. The learned A-Box primarily consists of the probabilities of each concepts. This is one area where we are eager to work on. Using the state-of-the art NLP techniques, we plan to fill this gap in a future work. Since our method has the potential to be used in any corpus, it could be seen that the lemmatizing and stemming algorithms that are available in WordNet would not recognize some of the words. Specially in the BioAssay corpus, we observe that some of the domain specific words are not recognized by WordNet. We use the Porter stemming algorithm [18] to get the word form and it shows that this algorithm constructs peculiar word forms. Therefore, we deliberately remove it from the processing pipeline.

The complexity of our algorithms is as follows. The bootstrapping algorithm available in the syntactic layer has a worst case running time of $O(M \times max(s_j) \times max(w_k))$, where M is the number of documents, s_j is a the number of sentences in a document, and w_k is the number of words in a sentence. The probabilistic reasoning algorithm has the worst case running time of $O(|\mathcal{L}| \times |SuperConcepts|)$, where $|\mathcal{L}|$ is the size of the lexicon and $|SuperConcepts|$ is the size of the super concepts set. The ontologies generated from the system are consistent with Pellet[8] and FaCT++[9] reasoners.

Finally, our method provides a process to create a lexico-semantic ontology for any domain. For our knowledge, this is a very first research on this line of

[8] http://clarkparsia.com/pellet

[9] http://owl.man.ac.uk/factplusplus/

work. So we continue our research along this line and to provide better results for future use.

5 Conclusion

We have introduced a novel process to generate an ontology for any random text corpus. We have shown that our process constructs a flexible ontology. It is also shown that in order to achieve high precision, it is paramount that the corpus should be large enough to extract important evidence. Our research has also shown that probabilistic reasoning on lexico-semantic structures is a powerful solution to overcome or at least mitigate the knowledge acquisition bottleneck. Our method also provides evidence to domain experts to build ontologies using a top-down approach. Though we have introduced a powerful technique to construct ontologies, we believe that there is a lot of work that can be done to improve the performance of our system. One of the areas our method lacks is the separation between concepts and individuals. We would like to use the generated ontology as a seed ontology to generate instances for the concepts and extract the individuals already classified as concepts. Finally, we would like to increase the lexicon of the system with more tags available from the Penn Treebank tag set. We believe that if we introduce more tags into the system, our system can be trained to construct human readable (friendly) concepts and relations names.

Acknowledgements

This work was partially funded by the NIH grant RC2 HG005668.

References

1. Banerjee, S., Pedersen, T.: The design, implementation and use of the n-gram statistics package. In: In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. pp. 370–381 (2003)
2. Bos, J., Markert, K.: Recognising textual entailment with logical inference. In: HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 628–635. Association for Computational Linguistics, Morristown, NJ, USA (2005)
3. Carlson, A., Betteridge, J., Wang, R.C., Hruschka, Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining. pp. 101–110. ACM, New York, NY, USA (2010)
4. Chemudugunta, C., Holloway, A., Smyth, P., Steyvers, M.: Modeling documents by combining semantic concepts with unsupervised statistical learning. In: ISWC '08: Proceedings of the 7th International Conference on The Semantic Web. pp. 229–244. Springer-Verlag, Berlin, Heidelberg (2008)
5. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence research 24, 305–339 (2005)

6. Cimiano, P., Völker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery (2005)
7. Clark, P., Harrison, P.: Large-scale extraction and use of knowledge from text. In: K-CAP '09: Proceedings of the fifth international conference on Knowledge capture. pp. 153–160. ACM, New York, NY, USA (2009)
8. Dellschaft, K., Staab, S.: Strategies for the evaluation of ontology learning. In: Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. pp. 253–272. IOS Press, Amsterdam, The Netherlands, The Netherlands (2008)
9. Ding, Z., Peng, Y.: A Probabilistic Extension to Ontology Language OWL. In: HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4. p. 40111.1. IEEE Computer Society, Washington, DC, USA (2004)
10. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)
11. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, Pearson Education International, 2. edn. (2009)
12. Koller, D., Levy, A., Pfeffer, A.: P-CLASSIC: A tractable probabilistic description logic. In: Proceedings of AAAI-97. pp. 390–397 (1997)
13. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering 7(4), 343–360 (2001)
14. Lukasiewicz, T.: Probabilistic description logics for the semantic web. Tech. rep., Nr. 1843-06-05, Institut fur Informationssysteme, Technische Universitat Wien (2007)
15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
16. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: the penn treebank. Comput. Linguist. 19(2), 313–330 (1993)
17. Poon, H., Domingos, P.: In: Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden. ACL (2010)
18. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
19. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edn. (2009)
20. Salloum, W.: A question answering system based on conceptual graph formalism. In: KAM '09: Proceedings of the 2009 Second International Symposium on Knowledge Acquisition and Modeling. pp. 383–386. IEEE Computer Society, Washington, DC, USA (2009)
21. SHOE: Example computer science department ontology, `http://www.cs.umd.edu/projects/plus/SHOE/cs.html`, last visited on June 14, 2010
22. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data and Knowledge Engineering 25(1-2), 161–197 (1998)
23. Tanev, H., Magnini, B.: Weakly supervised approaches for ontology population. In: Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. pp. 129–143. IOS Press, Amsterdam, The Netherlands (2008)
24. Völker, J., Hitzler, P., Cimiano, P.: Acquisition of owl dl axioms from lexical resources. In: ESWC '07: Proceedings of the 4th European conference on The Semantic Web. pp. 670–685. Springer-Verlag, Berlin, Heidelberg (2007)