PrOntoLearn: Unsupervised Lexico-Semantic Ontology Generation using Probabilistic Methods

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

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

Abstract. Formalizing an ontology for a domain is a tedious and cumbersome process. It is constrained by the knowledge acquisition bottleneck. There exist many text corpora that could be used to create ontologies. Here we provide a novel unsupervised bottom-up ontology generation method. This method is based on lexico-semantic structures and Bayesian reasoning to expedite the ontology generation process. We provide a quantitative and two qualitative results illistrating 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: probabilistic ontology learning

1 Introduction

An ontology is a formal, explicit specification of a shared conceptualization [11], [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 (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.

[4] http://bioassayontology.org/

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 5 provides a detail description of the experiments based on three different text corpora and the discussion. Finally, section 6 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 (first order logical reasoning).

Table 1. The summary of the related work. Probabilistic learning (PR), never ending language learning (NELL), discovery and aggregation of relations in text (DART), recognising 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 [10],[13],[15]	reasoning	available	available	prob. theory
NELL [3]	24×7 learning	fixed	dynamic	ML techniques
DART [8]	world knowledge	$\times$	$\times$	semi-automated
RTE [2], [14]	entailment	$\times$	$\times$	ATP
NLU [20]	commonsense rules	$\times$	$\times$	semi-supervised
Text2Onto [7]	ontology learning	$\surd$	$\surd$	semi-supervised
LexO [2]	complex classes	$\surd$	$\times$	semi-supervised
FCA [6]	taxonomy	$\surd$	$\times$	FCA
OP [5], [4], [23]	ontology population	available	available	semi-/supervised

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

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 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 built from a lexicon. A lexicon of a language is its vocabulary built from lexemes [12], [16]. A lexicon contains words belonging to a language and in our work individual words from the corpus will be treated as the vocabulary, thus, the lexicon of 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. The definition of the lexicon of our work is given as follows.

Definition 1. *A lexicon $\mathcal{L}_{\mathcal{O}}$ is the set that contains words belonging to the English vocabulary, which is part-of-speech (POS) type tagged with the Penn Treebank English POS tag set [17]. The set $\mathcal{L}_{\mathcal{O}}$ is built from the tag set: NN (noun, singular or mass), NNP (proper Noun, singular), NNS (noun, plural), NNPS (proper Noun, plural), JJ (adjective), JJR (adjective, comparative), JJS (adjective, superlative), VB (verb, base form), VBD (verb, past tense), VBG (verb, gerund or present participle), VBN (verb, past participle), VBP (verb, non-3rd person singular present), and VBZ (verb, 3rd person singular present)*

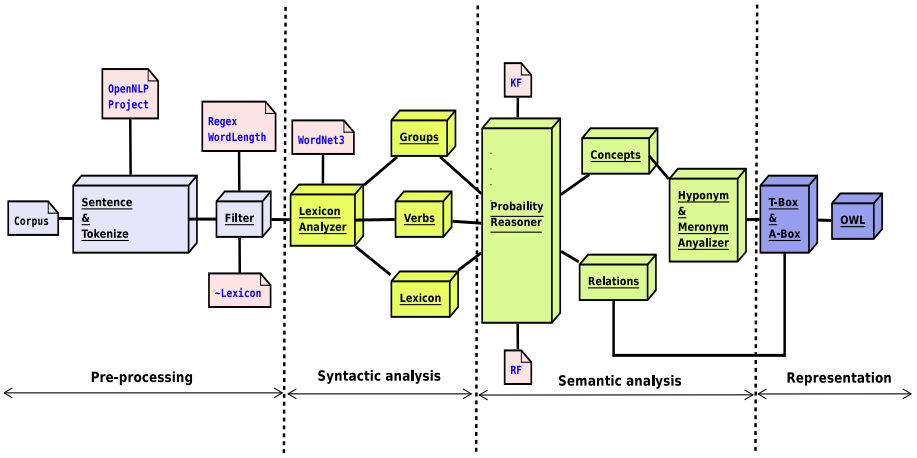


Fig. 1. Overall process: process categorizes into four phases; pre-processing, syntactic analysis, semantic analysis & representation

Definition 1 implies that negated lexicon $\overline{\mathcal{L}_{\mathcal{O}}}$ is the set that contains English words that are POS tagged with the Penn Treebank English POS tag set, other than the tags given in Definition 1. In addition, the word length W_L above some threshold W_{L_T} is also considered when building $\overline{\mathcal{L}_{\mathcal{O}}}$. The length of a word, with respect to POS context, is the sequence of characters or symbols that made up the word (e.g., the word "mika" has a word length of four $W_L = 4$). By default, we consider that a word with $W_L > 2$ sufficiently formalizes to some criterion.

Building up the pure lexicon at this stage, excluding the negated lexicon of the pre-processing, is known as tokenization from sentences [16]. Here, the pure form of the lexicon might contain words that need to be further purified according to some criterion. Words of the corpus contain many standard and constructed words. As mentioned, some words do not provide useful information (e.g., on, off, and at). In order to filter out these words, in the next phase of the pre-processing, each word is processed through a regular expression filter. The regular expression filter is a parameter of the system. The default regular expression is given as $[a-zA-Z]+$ if this parameter is not specified by the user. e.g., a word such as

du-145 will be filtered out from this regular expression. We also try to do token normalization to some extent. This is the process of canonicalizing the tokens so that matches occur despite superficial differences in the character sequences of the tokens [16]. In the next step, the vocabulary learned from the corpus is subjected to case-folding by reducing all letters to lower case. e.g., *Protocol* case-folds to *protocol*. Generally, documents use different forms of a word such as *organize*, *organizes* and *organizing* for grammatical reasons. 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 [16]. We achieve this with the aid of WordNets' stemming algorithms. We couple the knowledge of POS tag of the lexicon to get the correct context of the word.

3.2 Syntactic Analysis

The pre-processing phase eliminates the noise of the corpus and tags the $\mathcal{L}_{\mathcal{O}}$ according to Definition 1. The primary focus on this phase is to look at the structure of the sentences and learn the associations among the words in $\mathcal{L}_{\mathcal{O}}$. We assume that each sentence of the corpus follows the POS pattern in expression 1,

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

We hypothesize that the associations learned from this phase of the $\mathcal{L}_{\mathcal{O}}$ provides the potential candidates for concepts and relations of the ontology. But the words in the $\mathcal{L}_{\mathcal{O}}$ itself do 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.

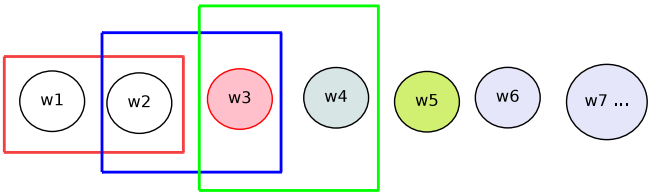


Fig. 2. An example three-gram model

According to Figure 2, 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 of $\mathcal{L}_{\mathcal{O}}$. 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. In addition to this, and in order to provide a useful taxonomy, we induce concepts from clustered concepts. Our defintion of concept learning is given in Definition 2.

Definition 2. *The set* $W = \{w_1, \ldots, w_n\}$ *represents independent words of the* $\mathcal{L}_{\mathcal{O}}$ *and each* w_i *has a prior probability* θ_i*. The set* $G = \{g_1, \ldots, g_m\}$ *represents independent 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. 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 *as shown in the Figure 3 and,*

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

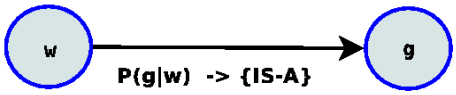


Fig. 3. Probabilistic IS-A relationship representation of the conceptualization (2). w and g are defined as the concepts of the conceptualization.

Lets define the *knowledge factor*: the lower-bound that select the super-concepts of the conceptualization.

Definition 3. $W = \{w_1, \ldots, w_n\}$ *represents independent words of the* $\mathcal{L}_{\mathcal{O}}$ *and each* w_i *has a prior probability* θ_i*. Lets define the knowledge factor (KF) as the lower-bound: if* $\theta_i \geq \tau$ *with* $0 \leq \tau \leq 1$ *then* w_i *is considered as a super-concept of the conceptualization.*

Definition 3 states that W of Definition 2 is considered as a super-concept of the conceptualization.

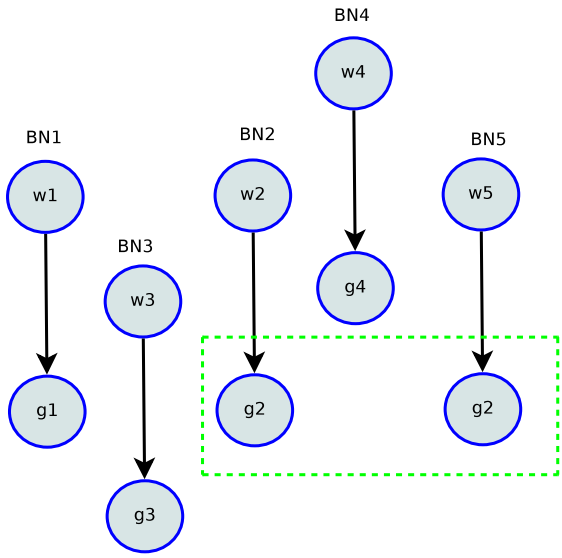


Fig. 4. w_1, w_2, w_3, w_4 and w_5 are super-concepts. g_1, g_2, g_3 and g_4 are candidate sub-concepts. There are 5 independent Bayesian networks. Bayesian networks 2 and 5 share the group g_2 when representing the concepts of the conceptualization

Definition 4. *The probabilistic conceptualization of the domain is represented by an n-number of independent Bayesian networks sharing groups (Figure 4).*

Figure 4 shows multiple Bayesian networks that share a common group g_2. The interpretation of Definition 4 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.

Definition 5. *Given a subset of concepts $G_S = \{g_1, \ldots, g_n\}$, $G_S \subset G$, with size n, for a given super-concpet w, when $P(g_1|w), \ldots, P(g_n|w)$ holds, the prefixes of the concepts are extracted and known as an induced concepts. For a m-gram model, at most up to $m-1$ concepts can be induced. For all induced concepts c, the concepts name collision will be avoided by assigning different namespaces. The induced concept will be given a prior probability of* 0.

Definition 5 gives an efficient way to represent the taxonomy of the conceptualization. Newly induced concepts contain words up to at most $m-1$. These concepts induction lead to concepts collision in the given namespace. This situation is avoided according to Defintion 6.

Definition 6. *When a concept is induced from a group of concepts, the induced concept is assigned to a different namespace in order to avoid possible concept name conflicts. The namespace assignment is forced, if and only if there exist a concept with the same name in the system, otherwise induced concepts will be subjected to the default namespace of the system.*

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*. The relations are as important as the concepts in a conceptualization. The relations exist among the concept of the conceptualization. We hypothesize that relations are generated by the verbs in the corpus.

Definition 7. *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.*

Definition 8. *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,*

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

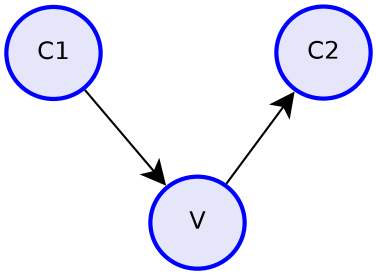


Fig. 5. Bayesian networks for relations modeling. C_1 and C_2 are groups and V is a verb

Using definitions 7 and 8, the relationship among multiple concepts are defined in 9. We define the relations in terms of groups of words in $\mathcal{L}_{\mathcal{O}}$. These groups are clustered around the most probable words found in the corpus.

Definition 9. *Let $S_p \subset S$ be a part of co-occurance sentence of the corpus, which can be transformed into $\{G_i\ v_j\ G_k\}$ groups and a verb. The sizes of G_i and G_k are $|G_i|$, $|G_k|$ and $G_i = \{g_1, \ldots, g_m\}$ and $G_k = \{g_{m+1}, \ldots, g_n\}$, $n > m$. Then, the relationships among G_i and G_k are build from the combinations of the elements from G_i and G_k with respect to v_j in accordance with the Bayesian model $p(G_i, G_k|V_j)$. There will be $|G_i| \times |G_k|$ relations,*

$$
\begin{aligned}
v_j(g_1, g_{m+1}) &\leftarrow p(g_1, g_{m+1}|v_j) \\
v_j(g_1, g_{m+2}) &\leftarrow p(g_1, g_{m+2}|v_j) \\
&\ldots \\
v_j(g_1, g_n) &\leftarrow p(g_1, g_n|v_j) \\
v_j(g_2, g_{m+1}) &\leftarrow p(g_2, g_{m+1}|v_j) \\
&\ldots \\
v_j(g_p, g_{m+1}) &\leftarrow p(g_p, g_{m+1}|v_j) \\
&\ldots \\
v_j(g_m, g_n) &\leftarrow p(g_n, g_m|v_j)
\end{aligned}
$$

The relations learned from Defintions 7 and 8 sometimes needs to be subjected to a lower bound. The lower bound is known as the *relations factor*, and it is used as an input parameter to the semantic analysis phase to set this lower bound.

Definition 10. *Let set $R = \{v_1(C_1, C_2), \ldots, v_m(C_k, C_r)\}$ be the relations that are learned from the corpus. Relations $v_i(C_j, C_k)$ are assigned a probability using a Bayesian model $p(C_j, C_k|V_i)$. When these relations are ordered based on their probability, a threshold φ is defined as the Relations Factor (RF) of the system.*

Definition 10 allows the user to limit the number of relations learned from the system. When the corpus is substantially large, the number of relations is proportional to the number of verbs in $\mathcal{L}_\mathcal{O}$. Not all relations may relevant and the RF is used as the limiting factor.

Definition 11. *Let v_i be a verb and v_j is the antonym verb of v_j learned from WordNet ($v_i \bowtie v_j$). Let there be relations $v_i(G_m, G_n)$ and $v_j(G_m, G_n)$ modeled by $p(G_m, G_n|v_r)$ ($v_r = i, j$). Since, $v_i \bowtie v_j$ for G_m and G_n, the relationship with the highest $p(G_m, G_n|v_r)$ value will be selected and the other relationship will be removed.*

We use verbs in $\mathcal{L}_\mathcal{O}$ as the key elements in forming relationships amoung concepts. Verbs have opposite verbs. Thus, according to Definition 11, if a verb is associated with some concepts and these concepts happen to be associated with a opposite verb, the verb with the highest Bayesian probability value is selected for the relations map and the other relationship will be removed from the system. Finally, the probabilistic conceptualization is serialized as an OWL DL ontology in the representation phase.

4 Implementation

The implementation of our approach uses a several open source projects to populate the required contexts at different phases as we introduce in section 3. The bootstrapping algorithm requires tokenizing sentences and stemming or lemmatizing of tokens to produce $\mathcal{L}_{\mathcal{O}}$ of the corpus. According to Defintion 1, $\mathcal{L}_{\mathcal{O}}$ is defined based on the Penn Treebank English POS tag set. We use the Stanford log-linear POS tagger[7], which uses the standard Penn Treebank tag set. We use the OpenNLP[8] project to produce sentences and tokens and the WordNet project to lookup for the type, stem and lemma of a word. In order to access the WordNet electronic library, we use the JWI[9] project. The BioAssayOntology corpus contains XHTML documents. We use the HTML parser[10] library to extract text from these documents. One of our other corpora contains PDF documents. We use the Apache PDFBox[11] library to extract the contents from the PDF documents. Finally, we use the Jena API[12] to serialize the probabilistic conceptualization model into OWL DL. Our implementation is based on Java 6 and it is named as `PrOntoLearn` (Probabilistic Ontology Learning).

5 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[13] domain . We have constructed ontologies for all three corpora with different parameter settings. One of the key problems we have encountered is the ontology evaluation. The BioAssay ontology dataset and the PDF dataset was impossible to evaluate as there are no existing reference ontologies or no ground truth that we could find of. Therefore, we use the third dataset from the University of Miami, Department of Computer Science domain and we conduct recall and precision given a reference ontology.

The first corpus, which is the primary data corpus of the experiment, contains high throughput screening assays performed on various screening centres. 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 from *[a-zA-Z]+[-_]?\w** regular expression, and normalized them to create $\mathcal{L}_{\mathcal{O}}$ of the corpus.

[7] http://nlp.stanford.edu/software/tagger.shtml
[8] http://opennlp.sourceforge.net/
[9] http://projects.csail.mit.edu/jwi/
[10] http://htmlparser.sourceforge.net/samples.html
[11] http://pdfbox.apache.org/
[12] http://jena.sourceforge.net/
[13] http://www.cs.miami.edu

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

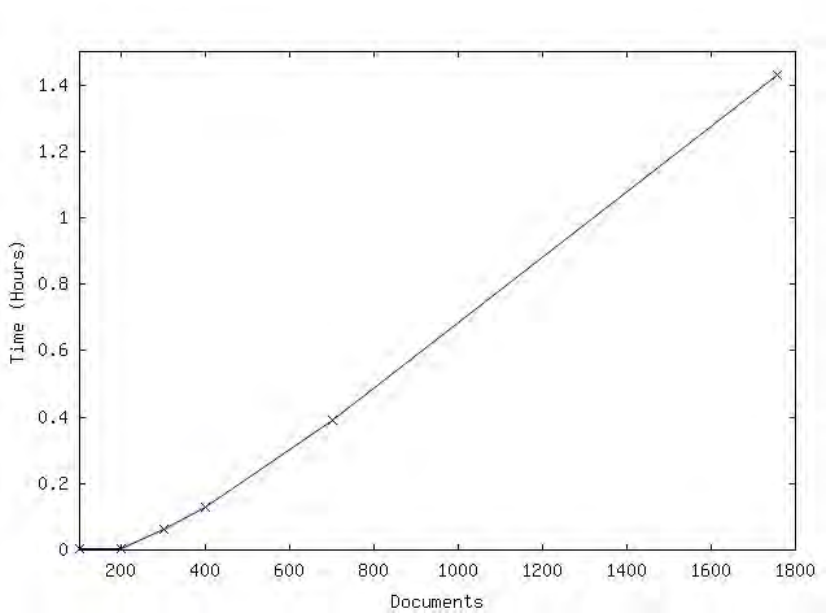
Title	Statistics	Description
Documents	1,759	All documents are XHTML formated with a given template
Unique $ConceptWords$	13,017	Normalized candidate concept words from NN, NNP, NNS, JJ, JJR & JJS using [a-zA-Z]+[-_]?\w*
Unique $Verbs$	1,337	Normalized verbs from VB, VBD, VBG, VBN, VBP & VBZ using [a-zA-Z]+[-_]?\w*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

Figure 6 shows the running time of the PrOntoLearn to build the probabilistic conceptualization model. It is found from the experiments that the POS tagger requires approximately 2,600 ms to train. 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. Figure 6 shows that the time required to build the conceptualization grows linearly.

One of the other obstacles we have encountered in terms of time complexity is in the representation layer. We use the Jena API to serialize the probabilistic conceptualization into OWL DL. When the system produced more than 1,000 concepts and relations, it is found that the Jena API takes a considerable amount of time to serialize the model. We use different architectural schemes to improve its performance. With all optimization, the presentation layer requires approximately 3.2 hours to serialize the model for the BioAssay Ontology corpus with full data set of 1,758 documents with capacity of 11.5 Mb. In order to provide a fast visualization of the conceptualization, we have written a simple yet flexible Java swing graphical user interface (GUI). This GUI has provided us visualizing and debugging the code as smoothly as possible. One of the other advantages of using a GUI is that it also provides the probabilities of the joint probability distribution $P(X, G)$, which is the representation of our probabilistic conceptualization.

The idea of our work is to generate an ontology without the supervision of a domain expert (unsupervised) for any given corpus. The user has to set system parameters such as KF, RF and regular expression of $\mathcal{L}_{\mathcal{O}}$. Since we use corpora from the bio medical domain, a collection of research papers and set of documents collected from computer science web site, the evaluation of the created ontology using standard techniques such as precision and recall is

Fig. 6. The BioAssay Ontology corpus vs. build time



not easy. We evaluate the generated ontologies with human domain experts. We obtain the comments and recommendations from the domain expert on the importance of the generated 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 7 and 8 show snapshots of the ontology created from the BioAssay Ontology corpus for input parametes $KF = 0.5$, N-gram = 3, and $RF = 0.9$. Figure 7 shows the IS-A relationships and Figure 8 shows the binary relationships.

We use a qualitative method to evaluate the BioAssay ontology using a human expert. According to the expert, the ontology contains rich set of vocabulary, which is very useful for top-down ontology construction. But expert also mentioned that the ontology have a flat structure. The main reason for this observation is that we use a 3-gram generator to create the ontology. Therefore, the maximum levels this model achieve is at most 3.

The *www.cs.miami.edu* corpus is used to calculate quantitative measurements. The gold standard based approaches such as precision (P), recall (R) and F-measure (F_1) are used to evaluate ontologies [9]. 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.

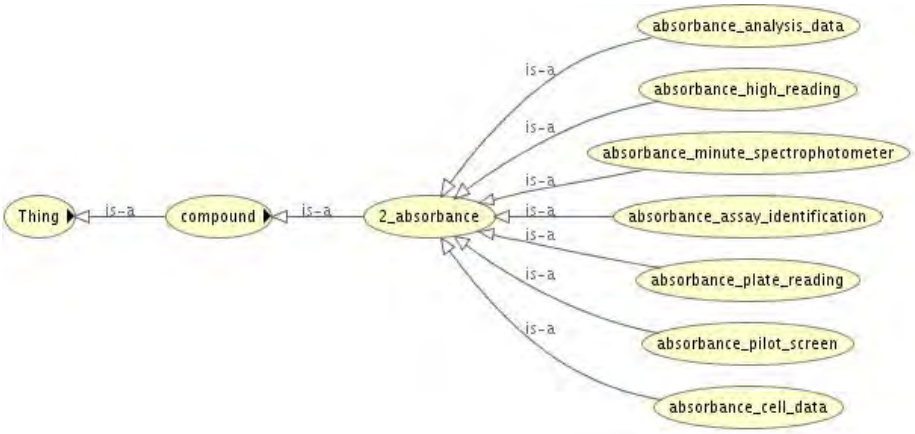


Fig. 7. An example snapshot of the BioAssay Ontology corpus with IS-A relations

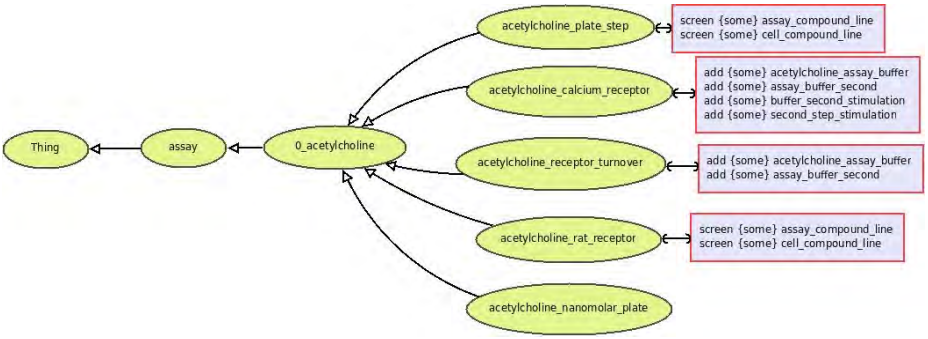


Fig. 8. An example snapshot of the BioAssay Ontology corpus with binary relations

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

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 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[14] and FaCT++[15] reasoners.

[14] http://clarkparsia.com/pellet
[15] http://owl.man.ac.uk/factplusplus/

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

6 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. We would use NLP technique to obtain this classification. In addition to this, our system can improve the quality of the relations if we introduce more specific grammar rules to sentences. We are looking at computational lexical semantics to prune the search space, so that the algorithms are efficient. 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.

References

1. Banerjee, S., Pedersen, T.: The design, implementation and use of the ngram 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.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
6. 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)
7. Cimiano, P., Völker, J.: Text2onto - a framework for ontology learning and data-driven change discovery (2005)
8. 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)
9. 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)
10. 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)
11. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. 5(2), 199–220 (1993)
12. 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. ed., [pearson international edition] edn. (2009)
13. Koller, D., Levy, A., Pfeffer, A.: P-classic: A tractable probabilistic description logic. In: In Proceedings of AAAI-97. pp. 390–397 (1997)
14. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Nat. Lang. Eng. 7(4), 343–360 (2001)
15. Lukasiewicz, T.: Probabilistic description logics for the semantic web. Tech. rep., Nr. 1843-06-05, Institut fur Informationssysteme, Technische Universitat Wien (2007)
16. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
17. 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)
18. Porter, M.F.: An algorithm for suffix stripping pp. 313–316 (1997)
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
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, The Netherlands (2008)