Web Intelligence and Agent Systems: An International Journal 8 (2010) 37–51
DOI 10.3233/WIA-2010-0178
IOS Press

Learning subsumption hierarchies of ontology concepts from texts

Elias Zavitsanos [a,b,*], Georgios Paliouras [a], George A. Vouros [b] and Sergios Petridis [a]
[a] *Institute of Informatics and Telecommunications, NCSR "Demokritos", Patriarhou Grigoriou and Neapoleos St., 15310, Athens, Greece*
E-mail: {izavits,paliourg,petridis}@iit.demokritos.gr
[b] *Department of Information and Communication Systems Engineering, University of Aegean, AI-Lab, 83200 Karlovassi, Samos, Greece*
E-mail: georgev@aegean.gr

Abstract. This paper proposes a method for learning ontologies given a corpus of text documents. The method identifies concepts in documents and organizes them into a subsumption hierarchy, without presupposing the existence of a seed ontology. The method uncovers latent topics for generating document text. The discovered topics form the concepts of the new ontology. Concept discovery is done in a language neutral way, using probabilistic space reduction techniques over the original term space of the corpus. Furthermore, the proposed method constructs a subsumption hierarchy of the concepts by performing conditional independence tests among pairs of latent topics, given a third one. The paper provides experimental results on the Genia and the Lonely Planet corpora from the domains of molecular biology and tourism respectively.

Keywords: Ontology learning, concept discovery, subsumption hierarchy construction, latent Dirichlet allocation, conditional independence

1. Introduction

Ontologies have been proposed as the key element to shape, manage and further process knowledge. However, the engineering of ontologies is a costly, time-consuming and error-prone task when done manually. Furthermore, in quickly evolving domains of knowledge, or in cases where information is constantly being updated, possibly making prior knowledge obsolete, the continuous maintenance and evolution of ontologies are tasks that require significant human effort. Thus, there is a strong need to automate the ontology development/maintenance tasks in order to minimize the cost of ontology creation and evolution.

For this reason, ontology learning has emerged as a field of research, aiming to help knowledge engineers to build and further extend ontologies with the help of automated or semi-automated machine learning techniques, exploiting several sources of information. Ontology learning is commonly viewed [1,10,30,35] as the task of *extending* or *enriching* an existing ontology with new ontology elements mined from text corpora. Depending on the ontology elements being discovered, existing approaches deal with the identification of concepts, subsumption relations among concepts, instances of concepts, or concept properties/relations. Linguistic, statistical, or machine learning techniques are used for these tasks.

The *seed* ontology used in ontology enrichment is usually a hierarchical backbone of concepts, related via subsumption relations, or a generic ontology that formalizes some of the concepts in a document collection. Linguistic approaches additionally suffer from language dependence, as they rely on language-specific lexico-syntactic patterns.

In contrast to the majority of the existing work, this paper proposes an automated approach to ontol-

*Corresponding author.

ogy learning, without presupposing the existence of a seed ontology, or any other type of external resource, except the corpus of training text documents. The proposed method addresses both tasks of concept identification and subsumption hierarchy construction. More specifically, concepts are identified and represented as multinomial distributions over the term space of the corpus. Towards this objective, the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [17] is used, following the Latent Dirichlet Allocation (LDA) [4] model. The discovered concepts, are then organized hierarchically by performing conditional independence tests among them. The statistical nature of the approach guarantees, among other, the language-independence of the proposed method. Finally, we extend our recent work [37] by presenting more extensive evaluation results, as well as by presenting a new gold standard-based evaluation method that takes into account the distributional representation of the learned topics, as well as a relative representation of the concepts of the gold ontology.

In what follows, Section 2 states the problem, refers to existing approaches that are related to the proposed method, and motivates our approach. Section 3 provides some backround knowledge concerning probabilistic topic models and the LDA model. Section 4 describes the proposed method, while Section 5 presents the evaluation method used to judge the performance of the method. Evaluation results are presented in Section 6, and finally, Section 7 concludes the paper by pointing out the advantages and limitations of the proposed method, sketching plans for future work.

2. Problem definition and related work

2.1. Problem definition

An ontology is a formal specification of a conceptualization of a domain, comprising concepts, individuals and properties. Ontology learning tries to learn automatically ontology elements and integrate them in an ontology, if one exists, in a consistent and coherent way. In this paper, we concentrate on the tasks of concept identification and taxonomy construction in the absence of prior knowledge, such as a seed ontology. Specifically, we deal with (a) the discovery of concepts from a given collection of documents, and (b) the hierarchical ordering of these concepts by means of the subsumption relation. Moreover, since (a) no prior knowledge is exploited and (b) only statistical and machine learning techniques are used, this paper aims to answer the following questions:

1. Is it possible to discover the concepts that express the content of documents in the corpus, independently of the terms' surface appearance?
2. Is it possible to form the ontology subsumption hierarchy backbone, using only statistical information concerning the discovered concepts?
3. Is it possible to devise a language-neutral ontology learning method?

Towards the identification of concepts and the learning of subsumption relations many approaches have been proposed. In the following subsection we describe the major ones that make use of linguistic, statistical and machine learning methods.

2.2. Concept identification

Starting with linguistic techniques for concept identification, the work in [23] uses pattern matching to derive noun phrases that indicate possible concepts. These approaches are based on matching regular expressions with Part-Of-Speech (POS) tags, in order to mark the desired noun phrases that follow a specific pattern. After tagging the texts, they extract units as candidate terms which take the form $((A \mid N)^+ \mid ((A \mid N)*(NP)^?)(A \mid N)*)N$, where A stands for adjective and N for noun, and their frequency of appearance is higher than a predefined threshold. The method of Moldovan and Girju [26] follows similar principles.

In addition, the morphology of words can be exploited in order to identify domain-specific terms. Small domain-specific units, e.g. morphemes or suffices, can indicate terms related to the domain of interest. Thus, the key idea is to identify useful character n-grams or morphemes and use them to select potential terms from the texts. Efforts like [21] and [8] have shown that the morphology of words can give important clues about their term status.

The use of prior knowledge is also helpful in the task of linguistic concept identification. When a word appears frequently together with a known term, they may constitute a new "complex" term. Moreover, if a word frequently appears together with some known terms in some specific pattern, the word becomes part of the terminology [9]. On this basis, the seed ontology provides the list of known terms, which are the lexicalizations of the concepts of the ontology.

Although linguistic approaches are widely adopted, they require significant text pre-processing and are

language-dependent. On the other hand, many statistical approaches to concept identification have also been proposed.

The authors of [10] apply statistical analysis to Web pages to identify words, which are then grouped into clusters that are proposed to the knowledge engineer. For this purpose, they make use of an existing ontology, the vocabulary as well as the relations of which, are exploited in order to construct a corpus by querying the WWW via Google. In this case, the ontology enrichment task is based on statistical information of word usage in the corpus and on similarity measures between concepts in the original ontology.

The authors in [2] extend an ontology with new concepts, taking into account words that co-occur with each of the existing concepts. The method requires that there are several occurrences of the concepts to be identified, so that there is enough contextual information to generate topic signatures. Topic signatures are usually sets of related words with associated weights. The work reported in [1] follows similar research directions.

More sophisticated schemes include the use of TF/IDF weighting in a corpus of documents to construct feature vectors for feeding a Latent Semantic Indexing (LSI) [11] process. Through this technique, latent topics are revealed which are actually distributions over the words of the term space of the corpus. The work in [6] and [5] also uses the method of LSI to retrieve latent entities in very large textual collections, as well as relationships between them. Probabilistic Latent Semantic Indexing (PLSI) has also been used in the task of concept identification [22]. It extends LSI assuming that each document is a probability distribution over topics and each topic is a probability distribution over words.

While approaches based on term frequency assume that the surface appearance of terms in documents provides sufficient information for concept discovery, more complex schemes, such as PLSI, suffer from overfitting to the training corpus, involving a large number of parameters that need to be estimated [4].

In this paper, in the phase of concept identification, we aim to uncover latent topics in the corpus, emphasizing the generative process of documents. Furthermore, these latent topics, which are represented as probability distributions over terms, mediate knowledge on the documents' contents. This approach is based on the assumption that the topics represent ontology concepts. Towards this target, we improve on previous approaches aiming to avoid overfitting and large sets of parameters, by using the Latent Dirichlet Allocation model.

2.3. Subsumption hierarchy construction

Subsumption hierarchy construction deals with the task of arranging the concepts identified in the previous step, in a hierarchy according to the subsumption relations that hold between them. The actual goal in this task is to identify the subsumption relations that hold between the ontology concepts.

Linguistic approaches usually construct subsumption hierarchies using lexico-syntactic heuristic patterns. Hearst patterns [19] are the most widely used and they are of the form:

- NP *such as* NP, NP, ..., and NP
- *such* NP *as* NP, NP, ..., or NP
- NP, NP, ..., *and other* NP
- NP, *especially* NP, NP, ..., and NP
- NP *is a* NP

Let as assume the phrase "There were several vehicles, such as cars, bikes and trucks". By applying a Hearst pattern we conclude that "cars", "bikes" and "trucks" are "vehicles". That is, the class or concept "vehicle" subsumes the classes or concepts "cars", "bikes" and "trucks", or one could say that "car" *is-a* "vehicle". However, since the *is-a* relation is sometimes confused with the *instance-of* relation, we reffer to hierarchical relations among concepts as subsumption or inclusion relations.

Hearst's idea was successfully applied in [24], while the authors in [27], based on Hearst patterns, defined several heuristics, like NP *find in* NP *such as* LIST, where LIST is a list of noun phrases, in order to construct a taxonomy of concepts.

At the linguistic level still, one can assume that a term A is a hyponym of a term B if A has more tokens than B, all the tokens of B are present in A, and both terms have the same head. Three provisions are needed for this to hold [29]. First, if a term includes dashes and brackets, then they should be ignored and the term should be considered as if there were no dashes and brackets. Second, a comparison of the lemmatized versions of the terms is needed. Third, the head of the term is the rightmost non-symbol token (a word). These provisions, though, make the approach specialized to the English language.

Linguistic approaches typically suffer from low recall, especially the ones based on pattern matching. This is due to the fact that the patterns do not occur

frequently enough in texts. Thus, subsumption relations can be learned only if they are explicitly mentioned in the corpus. Furthermore, such techniques seem to identify relations that hold mostly between words, rather than between concepts.

Besides the linguistic approaches that mainly identify subsumption relations between concepts, there are methods that deal at the same time with the task of concept identification and taxonomy construction. Moving towards to such methods, mainly machine learning and statistical ones, an extension of PLSI, named Hierarchical Probabilistic Latent Semantic Analysis (HPLSA) has been used in [12], in order to acquire a hierarchy of concepts, which are usually called topics in such methods. Due to its strong relation to PLSI, the drawbacks of PLSI mentioned in the previous section are inherited by this method. Hierarchical Latent Semantic Analysis (HLSA) has been applied in [28] to introduce hierarchical dependencies among topics by exploiting the word co-occurrences. This approach actually computes relations among topics in terms of words in the topics: those that appear in more than one topic at a specific level are grouped together at a higher level.

Finally, a hierarchical extension to LDA is presented in [3], where a latent hierarchy of topics is inferred from data. Although the branching factor at each level of the hierarchy is automatically determined, each document in the corpus is modeled as a path from the root topic to a leaf. As a result, each topic subsumes only one specific topic (leaf) and its abstractions, an approach that seems less flexible than the original approach of LDA, in which each document is a mixture over all the latent topics that are inferred.

Towards overcoming the problems of linguistic techniques, and the surface appearance of words, on which many statistical techniques rely, we follow a purely probabilistic approach to subsumption relation discovery. The proposed approach does not depend on the language or the annotation of the corpus. Instead it uses conditional independence tests on the latent topics discovered iteratively, in order to identify subsumption relations. Each document is modeled as a mixture over all the specific topics (leaves), as well as over all their abstractions at the previous level, etc. It must also be pointed out that, given the latent topics, the proposed method may compute more than one subsumption hierarchies. This is an additional benefit in cases where the domain cannot be modelled by a single hierarchy.

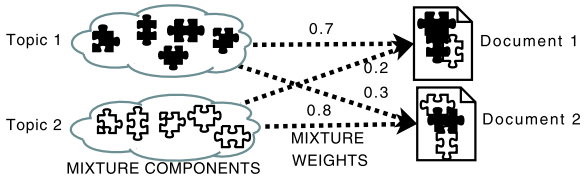

Fig. 1. The generative process: Documents are mixtures of topics. Topics are probability distributions over words (puzzle pieces). The probability of participation of a topic in a document is defined by the mixture weights. (Inspired from [33].)

3. Background on Probabilistic Topic Models

Probabilistic Topic Models (PTMs) [33] are based on the idea that documents are mixtures of topics, where a topic can be thematic and is represented by means of a probability distribution over words. PTMs follow the bag-of-words assumption, i.e. that words are independently and identically distributed in the texts. Topic models are generative models for documents: they specify a probabilistic procedure by which documents are generated. They are based on probabilistic sampling rules that describe how documents are generated as combinations of latent variables, i.e. the topics. Figure 1 illustrates the generative process: topics (clouds) are probability distributions over a predefined vocabulary of words (puzzle pieces). According to the probability that a topic participates to the content of each document, the process samples words from the corresponding topic in order to generate the documents.

In this paper, we are not interested in the generative process per se, but rather in the inverse process. Documents are known and words are observations towards assessing the topics of documents, as combinations of words. For this purpose, we use the LDA model described below.

The probabilistic generative process that is used in LDA states that topics are sampled repeatedly in each document. The generative process assumes the existence of K topics, providing some dependence from the surface appearance of terms. Specifically, given a predefined number of topics K, for each document:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
3. For each of the N words w_n:
   - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   - Choose a word w_n from $\text{p}(w_n \mid z_n, \beta)$, a multinomial probability distribution conditioned on the topic z_n.

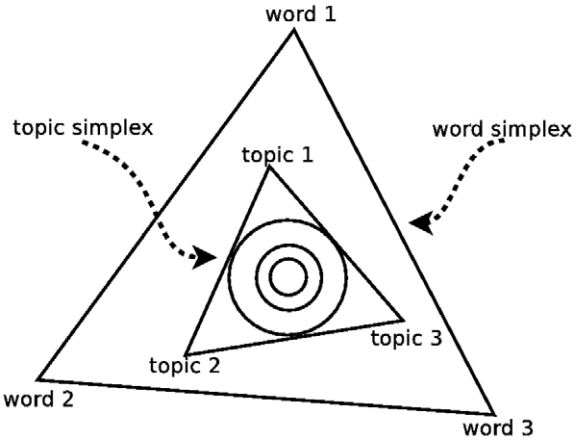

Fig. 2. The latent space formed by the LDA model for a vocabulary of $N = 3$ words and a predefined number of $K = 3$ latent topics (from [4]).

$p(z_n = i)$ stands for the probability that the i^{th} topic was sampled for the n^{th} word and indicates which topics are important, in the sense that they reflect the document's content for a particular document. $p(w_n \mid z_n = i)$ stands for the probability of the occurrence of word w_n given the topic i and indicates the probability of word occurrence for each topic. Thus, within a document, the probability distribution over words specified by the LDA model is given by Eq. (1).

$$p(w_n) = \sum_{i=1}^{K} p(w_n \mid z_n = i)p(z_n = i). \quad (1)$$

The basic assumptions of this model comprise the Poisson distribution that gives the length N of the documents, which is not critical in this process and thus, it can be replaced by a more realistic document length, and a Dirichlet prior on the multinomial distribution of topics. The Dirichlet prior is used to simplify the statistical inference, since it is the conjugate prior of the multinomial distribution. In addition, the number of topics K is assumed to be known and fixed. This parameter specifies the dimensionality of the Dirichlet distribution, and thus, the dimensionality of the topic variable z.

The topics that this probabilistic model generates form a latent space, where both documents and words can be represented. Figure 2 depicts the latent space for a vocabulary of $N = 3$ words and a predefined number of $K = 3$ latent topics.

The outer triangle, which is the word simplex, is the initial term space. Its corners correspond to the distributions where one of the three words has probability equal to one and the other two zero. The topics generated by LDA form a $(N - 1)$-dimensional simplex inside the initial space, which is the topic simplex. Thus, each topic is a specific distribution over words. Its corners correspond to those topic distributions where one of the three topics has probability equal to one and the other two zero. The documents that the generative process of this model creates, are placed inside the topic simplex over the contour lines of the Dirichlet distribution. Thus, documents are represented as mixtures of topics. The role of the Dirichlet parameter α is to determine how dominant a topic is going to be in a document. Low values of α will make one or two topics predominant in each document, whereas larger values will give similar weight to more topics. Therefore, changes to the parameter α lead to different placements of the inner topic simplex inside the word simplex. Concerning the number of topics (K), low values for K will result to a small number of "generic" topics, whereas larger values will result to a bigger number of more "specific" topics.

To sum up, through the LDA approach, the whole corpus is modeled as a Dirichlet parameter θ, governed by a prior α. The dimensionality of θ is equal to the number of topics that capture the knowledge of the domain covered by the text collection. The behavior of the K topics is determined by α.

As already pointed, in our case, where the objective is to discover concepts and order them in a subsumption hierarchy, the documents are known, and the observations are the terms appearing in the documents. So, we aim to infer the topics that generated the documents and then organize these topics hierarchically. The proposed method uses the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [17]. The reader is referred to [16] for a detailed explanation of this process.

4. The ontology learning method

4.1. Concept identification

As Fig. 3 illustrates, given a corpus of documents, the method first extracts terms by removing the stopwords from the texts. The extracted terms constitute the term space for the application of the LDA model described in Section 3. In the second step, feature vectors are constructed based on the document frequency of the terms in the documents. Next, the latent topics

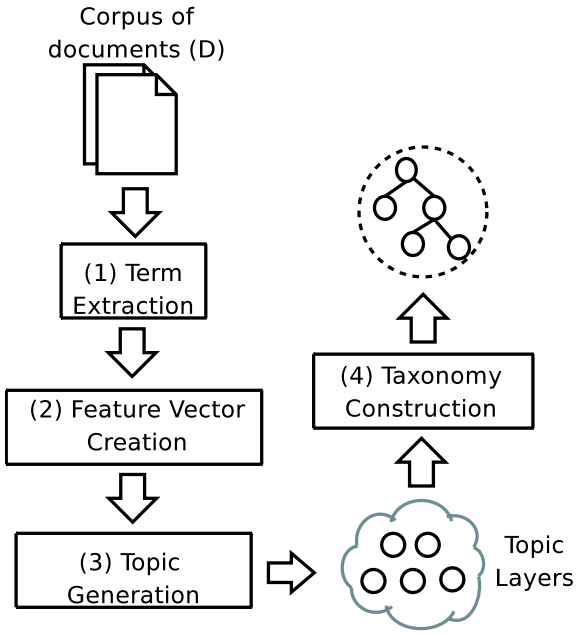

Fig. 3. The proposed method for ontology learning.

are generated as distributions over vocabulary terms according to the documents in the corpus and the terms observed in them. Finally, assuming that the topics generated correspond to ontology concepts, we organize them in a subsumption hierarchy according to their conditional independencies. In order to do so, we assume that a topic cannot subsume only one other topic, but it has to subsume at least two topics. At the end of the section we argue that having a topic that subsumes only one other topic leads to an incomplete modelling of the domain.

Therefore, the steps followed by the proposed method are as follows:

1. *Term Extraction* – From the initial corpus of documents, treating each document as a bag of words, we remove stop-words by creating a histogram of frequencies of the words that appear in the documents. This histogram follows a Zipfian distribution. Therefore by removing the most frequent words, we actually remove the majority of the stop-words. This technique is performed in order to maintain the language-independence of the method. However, standard lists of stop-words may also be used according to the language of the texts. The remaining words constitute the vocabulary and form the term space for the application of the LDA.
2. *Feature Vector Creation* – This step creates a Document – Term matrix of frequencies. Each cell of this matrix records the frequency of each term in each document of the corpus. This matrix is used as input to the LDA for topic generation.
3. *Topic Generation* – Sets of topics are generated at this step by the iterative application of the LDA for different values of the parameter L (number of topics). Therefore this step results in a multi-set of topics; each set being produced for a specific value of L. Starting from one topic at level l_0, the method iterates and terminates when L topic sets are produced. The topic set at each level l_i is larger than that of l_{i-1} by one topic. A small topic set forces a small number of topics to capture all the knowledge that the corpus contains, making the topics all-inclusive, and thus too generic in meaning. As the topic set increases iteratively, the generated topics become more focused, capturing more detailed domain knowledge. Thus, the method starts from "generic" topics, iterates and converges to more "specific" ones. In accordance to Fig. 2, at each iteration of the LDA, a new latent space of topics is created in the same term space. This is due to the fact that the dimensionality of the Dirichlet variable θ changes at each iteration, since the number of topics changes. Therefore, while a simple application of the LDA models the corpus as a unique Dirichlet variable θ, now the corpus is modeled through L Dirichlet variables, each for each level l_i, with different dimensionalities.

4.2. Taxonomy construction

Assuming that each of the computed topics corresponds to an ontology concept, the last step (Fig. 3) constructs the subsumption hierarchy of the discovered concepts. The concepts are arranged in a hierarchy according to their conditional independencies, given the topics at higher level. The intuition behind this is as follows: since the generated topics are random variables, e.g. *A* and *B* at level l_i, by measuring their mutual information we obtain an estimate of their mutual dependence. Therefore, given a third variable *C*, of the previous level, l_{i-1}, that reduces the mutual information of topics *A* and *B*, *C* contains a large part of the common information of *A* and *B*, i.e., *C* is a broader topic than the others. Topic *C* belongs in a topic set that contains broader in meaning topics than the ones in the set of *A* and *B*. In this case we may safely assume that

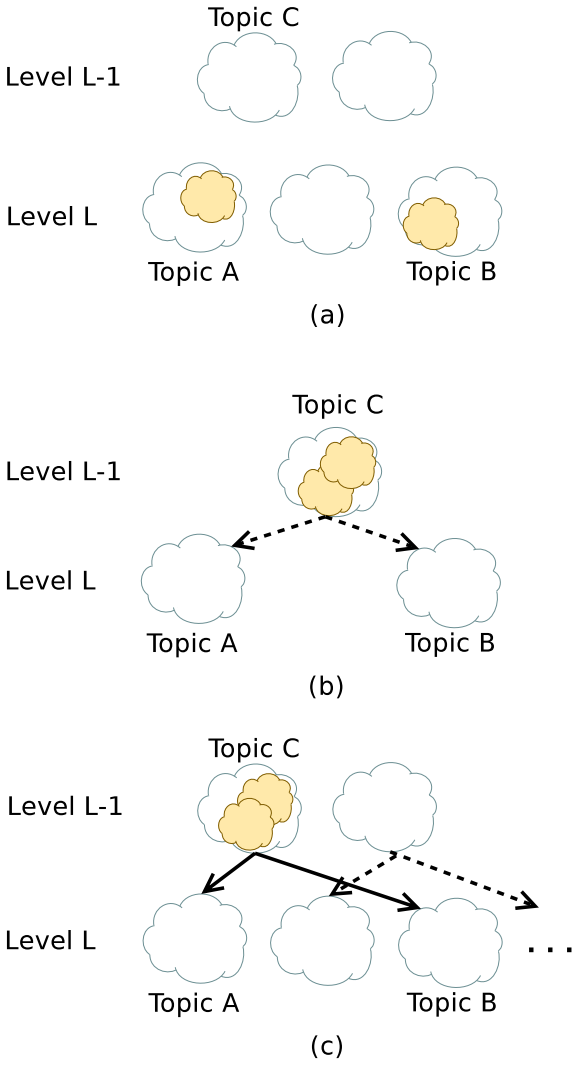

Fig. 4. The taxonomy construction process. Topics A and B have been generated in level l_i, while topic C in the previous level l_{i-1}. Topics A and B are mutually dependent given no prior knowledge. However, given topic C, the become conditionally independent (b). The broader topic C captures the mutual information of A and B, and thus the corresponding subsumption relations are added to the ontology (c). Topic C is the topic of level l_{i-1} that provides maximum independence between topics A and B. The process continues for other topic pairs in order to retrieve other subsumption relations.

C subsumes both *A* and *B*, and the corresponding relations are added to the ontology. Figure 4 depicts this process.

According to the iterative procedure of step 3, sets of "general" topics are being generated before the generation of sets of "specific" topics. In order to calculate the conditional independencies between topics, we take advantage of the document-topic matrix generated by the LDA model. Each entry of this matrix expresses the probability of a specific topic to participate in a specific document i.e., this is the probability of a topic, given a document. The process that generates the subsumption hierarchy is described by Algorithm 1.

Algorithm 1 Taxonomy construction using conditional independence tests.

for every topic set S_i **do**
  for every topic t_i in topic set S_i **do**
    for every pair of topics (t_j, t_k) in topic set S_{i+1} **do**
      if (conditional independence of t_j and t_k given t_i is the maximum among other pairs) AND (satisfies a threshold th) **then**
        t_i is parent of t_j and t_k
      end if
    end for
  end for
end for

Assuming that the topic sets have been generated through the iterative application of LDA, the algorithm starts from the first topic set that contains the most "general" topic and continues deeper in the hierarchy to larger topic sets. Given the set of topics at level l_{i+1}, the aim is to detect the pair of topics *(A,B)* whose independence is the maximum among the existing pairs of topics in l_{i+1}, given a topic *C* in l_i.

The conditional independence between two topics *A* and *B*, given a topic *C* is tested according to Eq. (2), where th is a threshold, having a very small value near zero (such as 10^{-7}) in order to avoid small rounding errors.

$$|P(A \cap B \mid C) - P(A \mid C)P(B \mid C)| \leqslant th \quad (2)$$

In order to compute Eq. (2), we need the probability of a topic *A* to participate in the corpus *D*, given that a topic *C* participates in the corpus. This is provided by Eq. (3).

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)}. \quad (3)$$

The probability of a topic *C* to participate in the corpus is given by the Eq. (4).

$$P(C) = \sum_{i=1}^{|D|} P(C \mid d_i)P(d_i), \quad (4)$$

where $|D|$ is the number of documents in the corpus and

$$P(d_i) = \frac{1}{|D|} \tag{5}$$

is the probability of a document in the corpus.

Accordingly, the joint probability of topics *A* and *B* to participate in the corpus, given that a topic *C* participates in the corpus, is given by Eq. (6).

$$P(A \cap B \mid C) = \frac{P(A \cap B \cap C)}{P(C)}. \tag{6}$$

Using the above equations, the mutual information between pairs of topics can be measured by Eq. (7).

$$I(A \cap B) = \sum_{a \in A} \sum_{b \in B} p(a,b) log \frac{p(a,b)}{p(a)p(b)}. \tag{7}$$

By maximizing the independence of two topics given a third one, we minimize their corresponding mutual information:

$$I(A \cap B \mid C) = \sum_{a,b \in A,B} \sum_{c \in C} p(a,b \mid c) log \frac{p(a,b \mid c)}{p(a \mid c)p(b \mid c)}, \tag{8}$$

since Eq. (9) holds due to Eq. (2).

$$log \frac{p(a,b \mid c)}{p(a \mid c)p(b \mid c)} = 0 \tag{9}$$

Ideally, one could set the threshold parameter th to zero and use the Eq. (10) to infer the conditional independencies between the topics. However, in such cases we assume that there are no rounding or other errors due to double precision concerning the computer numbering format.

$$|P(A \cap B \mid C) - P(A \mid C)P(B \mid C)| = 0 \tag{10}$$

Therefore, making such assumptions, Eq. (9) yields:

$$log \frac{p(a,b \mid c)}{p(a \mid c)p(b \mid c)} = log 1 = 0. \tag{11}$$

Since the algorithm searches for conditional independencies between pairs of topics in a topic set, it is not able to infer subsumption relations in the case where a topic subsumes only one other topic. However, this case of having a topic subsuming only one other topic would actually lead to an incomplete modeling of the domain. We would expect that since a concept *C* denotes a set of individuals, then a subsumee *A* of this concept would denote a subset of these individuals. Therefore, we would expect the existence of at least one more concept *B* that would denote individuals of *C*, that are not denoted by *A*.

Finally, although the maximum number of topics per level L affects the number of iterations of the algorithm, it must be pointed that it does not affect the depth of the produced hierarchy, leaving this choice to the algorithm. The depth of the hierarchy depends on the inclusion relations that are discovered between topics in different layers. Therefore, the number of topics only provides an upper limit to the depth of the learned ontology. For instance, one could set $L = 10$, forcing the algorithm to iterate 10 times in order to produce 10 topic sets, and thus 10 levels of the hierarchy. However, the algorithm may not infer any subsumption relation between the last two or three levels. Thus, the depth of the produced ontology would be 8 or 9, although we assumed that it should be 10.

5. Evaluation method

The corpora that we used are accompanied by the corresponding gold ontologies and we are interested in treating these ontologies as gold standards. Therefore, our evaluation method comprises the transformation of a gold ontology to a distributional representation, against which to evaluate the learned ontologies. In order to do that, we need to represent the gold standard ontology concepts in the same way as the learned topics, i.e. as multinomial probability distributions over the term space of the documents. Finally, a one-to-one matching of the gold concepts to the topics generated by the proposed method is performed to assess the quality of the learned ontology.

Concerning the transformation of the gold standard ontology in order to represent each concept as a distribution over terms, we measure the frequency of the terms that appear in the "context" of each concept. In both corpora that we used, the concept instances are annotated in the texts, providing direct population of the concepts in the golden standard ontologies with their instances. Therefore, as Fig. 5 illustrates, we perform the following transformation procedure:

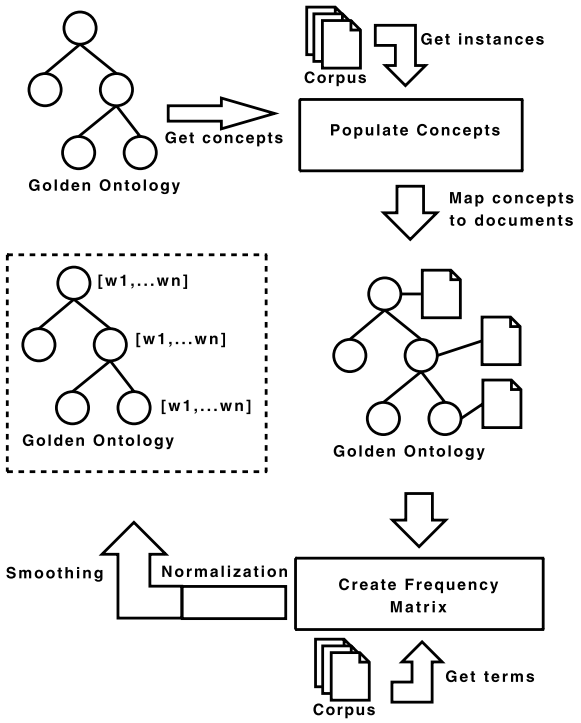

Fig. 5. The transformation of the gold standard ontology concepts into probability distributions. Concepts are first populated in order to locate their contexts. Then, vectors of term frequencies are created based on the context of each concept. Finally, normalization and smoothing is performed to transform these vectors into probability distributions.

1. Populate Concepts – By searching the document space, we retrieve all the concept instances that appear and we directly populate the gold ontology with instances.
2. Map Concepts to Documents – After the population of each concept with its instances, it is possible to associate each document to the concept(s) that it refers to. This is performed by counting the concept instances that appear in each document.
3. Create Frequency Vectors – Having the mapping between concepts and documents we create feature vectors based on the document in which each concept appears. These vectors have the form of a two-dimensional matrix that records the frequency of each term in the context of each concept. That is, we have a "concept – term" matrix that represents each concept as a distribution over the term space of the text collection. The context of each concept in this case is the whole document that is associated with this concept.
4. Normalization – For each concept, the frequencies are normalized giving a probability distribution over the term space. In addition, a smoothing of the probability distributions is performed to eliminate possible zero values of unseen terms, using Eq. (12), where N is the size of the term space.

$$\hat{P}_L(w_i) \doteq \frac{\hat{P}(w_i)+1}{N+1}, \forall i. \tag{12}$$

Using the new representation of the golden concepts, a one-to-one matching to the generated topics can be performed. Since both representations are based on probability distributions, we used the symmetric KL divergence (13) for matching.

$$D_{KL} = \frac{1}{2}\Bigg[\sum_i P(w_i) log \frac{P(w_i)}{Q(w_i)} + \sum_i Q(w_i) log \frac{Q(w_i)}{P(w_i)}\Bigg]. \tag{13}$$

For a golden concept p and a generated topic q, $P(\cdot)$ and $Q(\cdot)$ are the corresponding probability distributions. Small values of KL divergence indicate high similarity between a concept p and a topic q.

Therefore, a topic is matched to a concept if their corresponding distributions are the "closest" compared to all the other and their KL divergence is below a fixed threshold th_{KL}. The threshold th_{KL} affects the matching of topics to gold concepts in the sense that strict choices very close to zero cause few topics to be matched with gold concepts and loose choices lead to more matchings. Since small values of KL divergence indicate high similarity between a topic and a concept, ideally, a value equal to zero would indicate a perfect match. However, such a situation is unlikely to happen, since the topics are produced by the application of LDA and the gold concepts are transformed to distribution by the transformation method of this section. Therefore, we expect that we cannot have identical topics to concepts. Thus, we assume that values different from zero, but rather strict near 0.1 are appropriate for out evaluation method.

6. Experimental results

We have evaluated the proposed method, using the method of Section 5, on two corpora:

Table 1
Information regarding the datasets: the number of documents of each dataset, the number of concepts in the gold ontologies, the number of instances and relations in the gold ontologies, as well as the average number of instances per concept

Corpus	Documents	Gold concepts	Instances	Relations	Avg. instances/concept
Genia	2000	43	14000	41	300
LP	300	60	600	60	5

1. The Genia corpus, which contains 2000 documents from the domain of molecular biology and is accompanied by the Genia ontology. The ontology comprises 43 concepts connected by 41 subsumption relations, which is the only type of relation among the concepts. The Genia corpus contains about 14.000 concept instances in 2000 documents. In particular, the average number of instances per concept is about 300. These resources are available from the Genia project.[1]
2. The Lonely Planet corpus, which is a collection of about 300 Web pages from the Lonely Planet Web site,[2] providing touristic information. The corpus contains about 600 concept instances in these 300 Web pages, while the average number of instances per concept is about 5. The corresponding ontology contains 60 concepts and 60 subsumption relations among them, which is the only type of relation among them.

The corresponding ontologies of both corpora served as gold standards for evaluation. Table 1 summarizes the above information.

The computation of the latent topics was done with a stand-alone Java application, making use of the Gibbs sampling approximation method. The parameters involved are the maximum number of topics (L) and the threshold th_{KL} introduced in our evaluation method in order to match learned topics to golden concepts. Since Algorithm 1 performs an exhaustive search to find the best solution, its complexity is $O(L^3)$. Although there is still room for making this algorithm more efficient, the implemented algorithm needed only 4 minutes to compute the hierarchies of our experiments on a standard Pentium 3.0 GHz PC.

We have experimented in both corpora for various values of the parameter L, setting strict values near zero to the parameter th_{KL} that affects the matching of the generated topics and the golden concepts, during evaluation. The method was evaluated in terms of *Precision* and *Recall*. Regarding the concept identification task, we define *Precision* as the ratio of the number of concepts correctly detected to the total number of concepts detected, and *Recall* as the ratio of the number of concepts correctly detected to the number of concepts in the gold standard. Accordingly, for the subsumption hierarchy construction task, *Precision* is the ratio of the number of subsumption relations correctly detected to the total number of subsumption relations detected, and *Recall* is the ratio of the number of subsumption relations correctly detected to the number of subsumption relations in the gold standard.

The *F-measure* is a combined metric the reflects the harmonic mean between precision and recall, and is defined as follows:

$$Fmeasure = \frac{2 * precision * recall}{precision + recall}. \tag{14}$$

The threshold th_{KL} affects the results in the sense that strict choices very close to zero cause few topics to be matched with golden concepts and loose choices lead to more matchings. Therefore, we evaluated the proposed method choosing a rather strict value for the threshold th_{KL}. Figure 6 depicts how the F-measure is affected for two different values of threshold th_{KL}, concerning both concept identification and taxonomy construction in the case of the Genia corpus. In all diagrams, the X-axis ranges according to the number of topics inferred by the proposed method.

Accordingly, Fig. 7 depicts how the F-measure is affected for two different values of threshold th_{KL}, concerning both concept identification and taxonomy construction in the case of the Lonely Planet corpus.

For small values of th_{KL}, the results are worse than for larger ones. Actually, by setting such small values, we require the learning method to infer topic distributions that are almost identical to the ones that correspond to the gold standard ontology. On the other hand, since these distributions are over the whole term space of the dataset, we require that significant words in the topic distributions have analogous significance in the gold distributions.

[1] The Genia project, http://www.tsujii.is.s.u-tokyo.ac.jp/GENIA.
[2] The Lonely Planet travel advise and information, http://www.lonelyplanet.com.

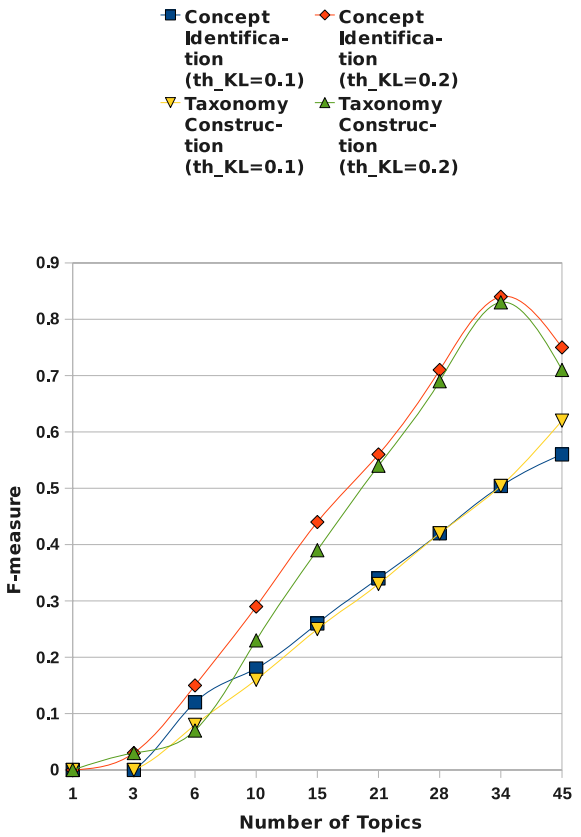

Fig. 6. F-measure for Concept Identification and Taxonomy Construction tasks for different values of the threshold th_{KL} for various numbers of topics in the case of the Genia corpus.

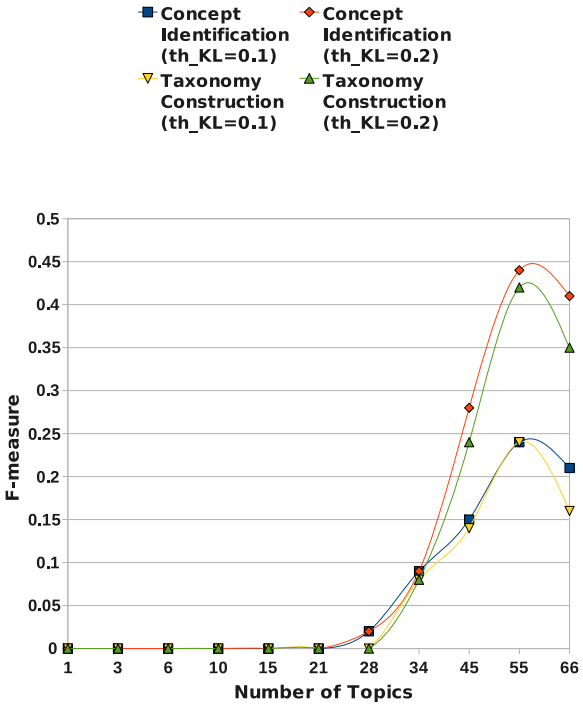

Fig. 7. F-measure for Concept Identification and Taxonomy Construction tasks for different values of the threshold th_{KL} for various numbers of topics in the case of the Lonely Planet corpus.

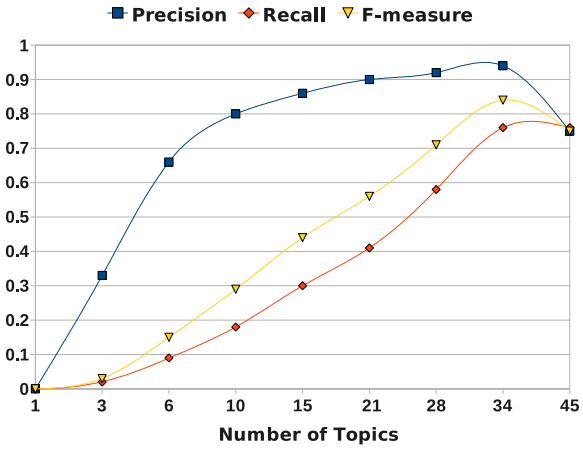

Fig. 8. Precision, Recall and F-measure for the task of concept identification for various numbers of topics in the case of the Genia corpus.

The fact that the results for both tasks of concept identification and taxonomy construction behave similarly is due to the high relation between these two tasks. Missing one concept, actually indicates a possible miss of the corresponding subsumption relation. Since we assume that a correctly retrieved subsumption relation is this between two correctly identified concepts, a failure on the task of concept identification also affects the task of taxonomy construction.

Regarding the evaluation of both tasks we chose a rather low threshold value of th_{KL} equal to 0.2 for both corpora. Figure 8 depicts in more detail the results obtained on the Genia corpus for the task of concept identification, while Fig. 9 depicts the results for the task of subsumption hierarchy construction on the same corpus. In the latter Figure, the number of the correct identified subsumption relations for small hierarchies of 3 or 6 topics is the same, which explains the non-monotonic behavior of precision between these two points.

In the case of Genia, by retrieving 34 topics, the proposed method managed to create an ontology very close to the gold standard. Specifically, for this number of topics and for the task of concept identification, the precision is equal to 0.94, while the recall remains also high, equal to 0.76, despite the fact that the golden

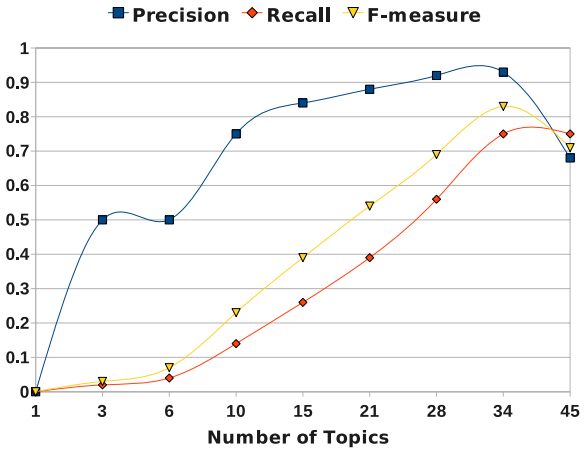

Fig. 9. Precision, Recall and F-measure for the task of subsumption hierarchy construction for various numbers of topics in the case of the Genia corpus.

ontology contains more concepts. Increasing L further, does not seem to improve the performance, since recall remains the same, while precision falls. This is due to the inability of the method to identify some very specific topics. Particularly, there are some very specific concepts subsumed by RNA and DNA that it is very hard to be distinguished by the learning method. We believe that these concepts are not concrete enough to be distinguished, and this is why the learning method clusters all their instances in one or two topics, forcing this way the evaluation method to penalize more the produced ontology.

In the task of taxonomy construction, precision is equal to 0.93, while recall is 0.75 for the same number of topics. For these values of precision and recall, Fig. 10 depicts a part of the learned taxonomy in comparison to the gold standard.

Concerning the Lonely Planet corpus, Fig. 11 provides quantitative results for the task of concept identification, while Fig. 12 presents the evaluation results for the task of subsumption hierarchy construction. Finally, Fig. 13 depicts a part of the learned taxonomy in comparison to the gold standard.

In the case of the Lonely Planet corpus the best results were achieved for 55 topics, which is in accordance to the fact that the gold ontology is larger than the one for Genia. For this number of topics and for the task of concept identification precision was equal to 0.62 and recall 0.36. Accordingly, for the task of taxonomy construction, the best quantitative results were achieved also for 55 topics and precision was 0.53, while recall was 0.35.

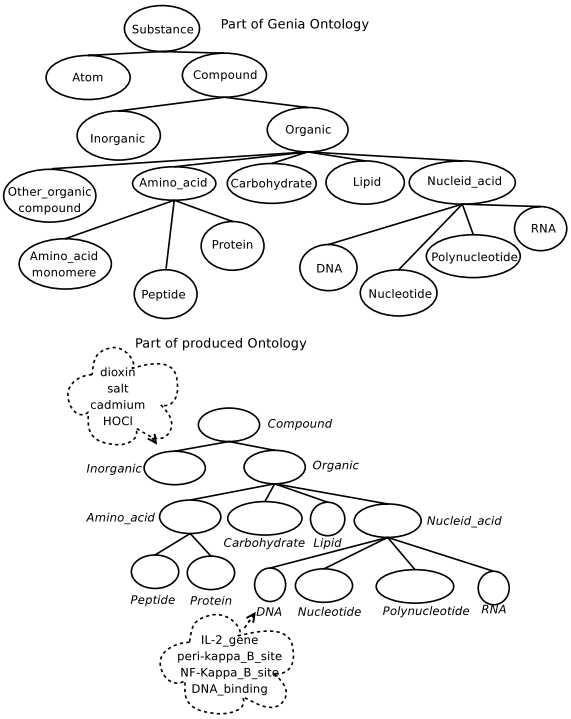

Fig. 10. Parts of the produced ontology and the GENIA ontology. In clouds: important terms that participate in the corresponding topics, which are also concept instances of the GENIA ontology.

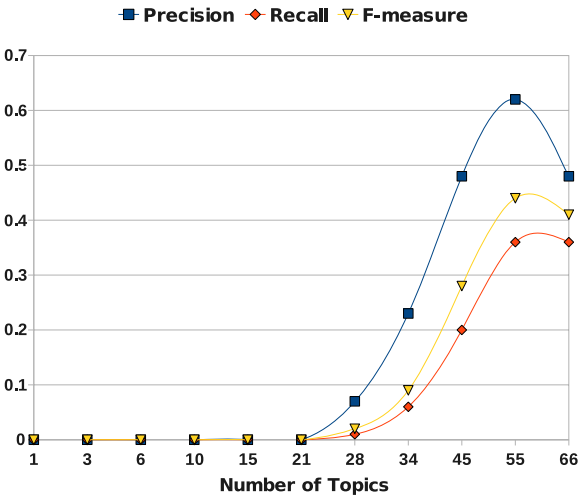

Fig. 11. Precision, Recall and F-measure for the task of concept identification for various numbers of topics in the case of the Lonely Planet corpus.

While in the case of the Genia corpus the results were very close to the golden standard, in the case of Lonely Planet, we observed that the learned ontology differs substantially from the gold standard. This result

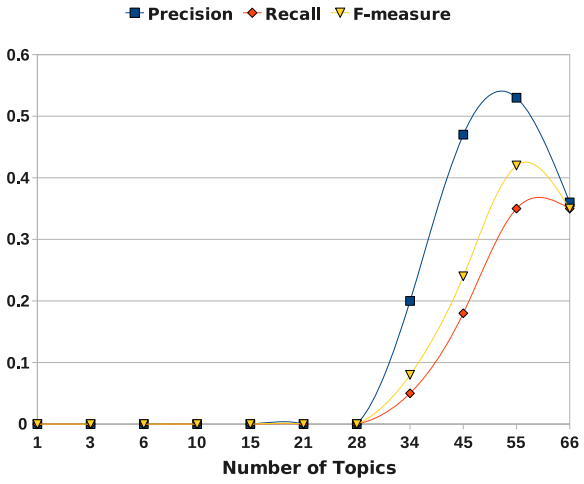

Fig. 12. Precision, Recall and F-measure for the task of subsumption hierarchy construction for various numbers of topics in the case of the Lonely Planet corpus.

is attributed to the fact that in the case of the Lonely Planet corpus, half of the golden concepts had a single instance and generally most of the concepts were insufficiently instantiated in the texts. Therefore, the difficulty of the model to discover some concepts explains the lower results compared to the Genia corpus. Although the concept instances that are annotated in the texts are not exploited directly by the learning method, having concepts that are instantiated sufficiently in the texts surely helps LDA to discover a more accurate model of the domain knowledge. For instance, spatial concepts, such as "Area", "City", "Country", "Island", and "Region", are among the ones that are instantiated frequently enough in texts, since the corpus deals with the tourism domain, and are among the ones that are correctly defined by the learning method. On the other hand, concepts like "Program", "Castle", and "FreeWay" were insufficiently instantiated and it was hard to be discovered.

Moreover, the fact that the model of LDA ignores topic correlations, in the sense that it is unable to model them due to the nature of its generative process, assuming that the produced topics are independent from each other, where they are not, introduces an additional difficulty to the discovery of a large number of fine-grained, tightly-coherent topics [36].

In addition, the choice of th_{KL} parameter for the evaluation task plays an important role on how we interpret the quantitative evaluation results. For instance, in Fig. 10, the inferred topic "Inorganic" among its most probable words contains all the instances of the corresponding gold concept. Setting a low value in

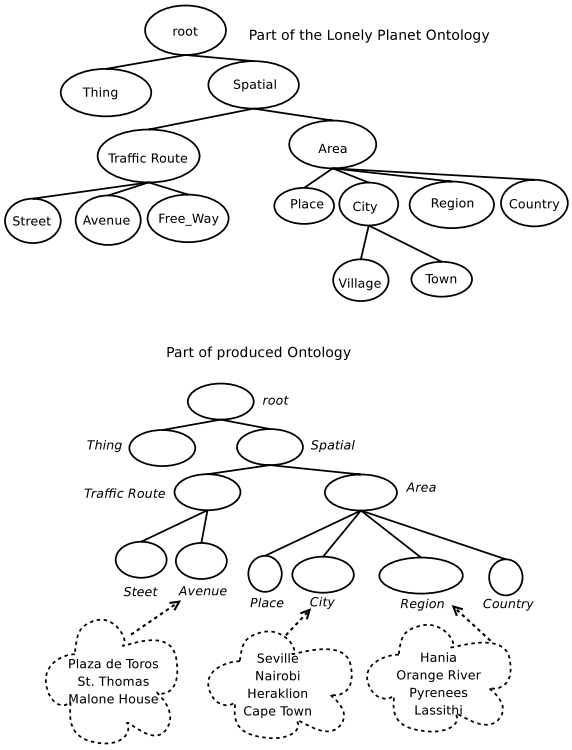

Fig. 13. Parts of the produced ontology and the Lonely Planet ontology. In clouds: important terms that participate in the corresponding topics, which are also concept instances of the Lonely Planet ontology.

th_{KL} leads to a heavy penalty since the evaluation takes into account all the words that participate in the probability distribution of this topic, and ignores the fact that the words that best describe this topic are in fact the correct instances of the corresponding gold concept. Therefore, in this evaluation scenario we conclude that values near 0.2 reflect the performance of the proposed method while at the same time the evaluation is quite strict. Finally, in the task of ontology learning is it necessary to provide qualitative results along with the quantitative ones in order to draw conclusions about the learning method.

Concluding this section, obviously a crucial parameter of the proposed method is the number of topics. Although the specific approach actually requires an upper bound to the the number of topics, since through the conditional independence tests discards topics that are not connected, it is important to be able to set rational values to this parameter a priori. In general, through this parameter, one sets the level of "specificity" of the learned hierarchy. However, in cases where no gold ontology is provided, finding the

optimal number of concepts in respect to the datasets requires cross-validation or the incorporation of methods, such as Hierarchical Dirichlet Distributions [34], that provide a prior over the number of topics.

Finally, the evaluation of ontologies when these ontologies are produced by an automated learning procedure is an open field of research. The research community has not established a standard methodology for automating ontology evaluation. Especially when the evaluation is done against a gold standard ontology, it seems that we cannot judge objectively the result, since the gold standard has been created by humans in a possibly subjective and biased manner. Particularly, in cases where the ontology has been learned from scratch and it is not the result of enrichment of a seed ontology, the evaluation is even more difficult.

7. Conclusions

In this paper we have proposed a fully-automated method for learning ontologies. The proposed method uses the Latent Dirichlet Allocation model for the discovery of topics that represent ontology concepts. According to this method, topics are represented as multinomial distributions over document terms. A method that performs conditional independence tests among topics arranges concepts in a subsumption hierarchy.

The major advantage of this approach is its statistical nature, which is based on probabilistic topic models. This allows the computation of topics in a language-neutral way, revealing those topics that express the contents of documents, and thus, the concepts that express the knowledge that documents mediate. This makes the method very generic, tackling at the same time both problems of concept identification and hierarchy construction.

In addition, a method for evaluating learned ontologies was presented and was used to obtain experimental results on the Genia and Lonely Planet ontologies and the associated corpora. The results that we obtained were very encouraging showing that the proposed method was able to reconstruct large parts of the golden ontologies only from the statistical analysis of the corpora. One weakness of the method that was identified in the experiments was the difficulty of discovering some very specific topics, especially in cases where they are not instantiated sufficiently in the texts.

Further work includes the improvement of our evaluation method towards semantic mapping [32] between the learned and the gold ontology, as well as further experimentation with more data sets and other (approximation) methods for inferring the latent topics.

Acknowledgments

The authors would like to acknowledge support by the research and development project ONTOSUM,[3] funded by the Greek General Secretariat for Research and Technology.

References

[1] Agirre, E., Ansa, O., Hovy, E., and Martinez, D., Enriching Very Large Ontologies Using the WWW, In ECAI 2000 Workshop on Ontology Construction, 2000.

[2] Alfonseca, E. and Manandhar, S., An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery, In Proceedings of the International Conference on General WordNet, 2002.

[3] Blei, D.M., Griffiths, T.L., Jordan, M.I., and Tenenbaum, J.B., Hierarchical Topic Models and the Nested Chinese Restaurant Process, In Advances in Neural Information Processing Systems, 2004.

[4] Blei, D.M., Ng, A.Y., and Jordan, M.I., Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, pp. 993–1022, 2003.

[5] Bradford, R.B., Efficient Discovery of New Information in Large Text Databases, IEEE International Conference on Intelligence and Security Informatics, LNCS, vol. 3495, pp. 374–380, 2005.

[6] Bradford, R.B., Relationship Discovery in Large Text Collections Using Latent Semantic Indexing, In Proceedings of the Fourth Workshop on Link Analysis, 2006.

[7] Buitelaar, P., Olejnic, D., and Sintek, M., A Protegé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis, In European Semantic Web Symposium, 2004.

[8] Cohen, J.D., Highlights: Language and Domain Independent Automatic Indexing Terms for Abstracting, Journal of the American Society for Information Science, 46(3), pp. 162–174, 1995.

[9] Enguehard, C. and Pantera, L., Automatic Natural Acquisition of A Terminology, Journal of Quantitative Linguistics, 2, pp. 27–32, 1994.

[10] Faatz, A. and Steinmetz, R., Ontology Enrichment with Texts from the WWW, In ECML/PKDD 2002 Semantic Web Mining Workshop, 2002.

[11] Fortuna, B., Mladevic, D., and Grobelnik, M., Visualization of Text Document Corpus, Informatica, 29, pp. 497–502, 2005.

[12] Gaussier, E., Goutte, C., Popat, K., and Chen, F., A Hierarchical Model for Clustering and Categorising Documents, LNCS, vol. 2291, pp. 229–247, 2002.

[13] GENIA project, http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA.

[3] See also http://www.ontosum.org/.

[14] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., Markov Chain Monte Carlo in Practice, London: Chapman and Hall, 1996.

[15] Gómez-Pérez, A. and Manzano-Macho, D., A Survey of Ontology Learning Methods and Techniques, In Ontology-based Information Exchange for Knowledge Management and Electronic Commerce, Deliverable 1.5.

[16] Griffiths, T. and Steyvers, M., A Probabilistic Approach to Semantic Representation, In Proceedings of the 24th Annual Conference of the Cognitive Science Society, 2002.

[17] Griffiths, T.L. and Steyvers, M., Finding Scientific Topics, In Proceedings of the National Academy of Science, 2004.

[18] Gruber, T.R., Toward Principles for the Design of Ontologies Used for Knowledge Sharing, In International Journal of Human Computer Studies, 1995.

[19] Hearst, M.A., Automatic Acquisition of Hyponyms from Large Text Corpora, In Proceedings of the International Conference on Computational Linguistics, 1992.

[20] Hearst, M.A., Automated Discovery of WordNet Relations, In WordNet: An Electronic Lexical Database, 1998.

[21] Heid, U., A Linguistic Bootstrapping Approach to the Extraction of Term Candidates from German Text, Terminology (Amsterdam), 5(2), pp. 161–182, 1999.

[22] Hofmann, T., Probabilistic Latent Semantic Indexing, In Proceedings of the 22nd annual international ACM SIGIR, 1999.

[23] Justeson, J.S. and Katz, S.M., Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text, Natural Language Engineering, 1(1), pp. 9–27, 1995.

[24] Kietz, J.U., Maedche, A., and Volz, R., A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet, In EKAW'00 Workshop on Ontologies and Texts, 2000.

[25] Maedche, A. and Staab, S., Ontology Learning, In Handbook on Ontologies in Information Systems, 2004.

[26] Moldovan, D.I. and Girju, R.C., An Interactive Tool for the Rapid Development of Knowledge Bases, Journal on Artificial Intelligence Tools, 10(1–2), pp. 65–86, 2001.

[27] Morin, E., Automatic Acquisition of Semantic Relations Between Terms from Technical Corpora, In Proceedings of the 5th International Congress on Terminology and Knowledge Engineering – TKE, 1999.

[28] Paaß, G., Kindermann, J., and Leopold, E., Learning Prototype Ontologies by Hierarchical Latent Semantic Analysis, In Knowledge Discovery and Ontologies (KDO-2004), 2004.

[29] Rinaldi, F. and Yuste, E., Exploiting Technical Terminology for Knowledge Management, In P. Buitelaar, P. Cimiano, B. Magnini (eds.), Ontology Learning and Population, IOS Press, 2005.

[30] Roux, C., Proux, D., Rechermann, F., and Julliard, L., An Ontology Enrichment Method for a Pragmatic Information Extraction System Gathering Data on Genetic Interactions, In ECAI 20000 Workshop on Ontology Learning, 2000.

[31] Salton, G. and McGill, M.H., Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.

[32] Silva, N. and Rocha, J., Semantic Web Complex Ontology Mapping, International Journal of Web Intelligence and Agent Systems, IOS Press, 1(3), pp. 235–248, 2003.

[33] Steyvers, M. and Griffiths, M., Probabilistic Topic Models, In Handbook of Latent Semantic Analysis, 2007.

[34] Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M., Hierarchical Dirichlet Processes, Journal of the American Statistical Association, 101, pp. 1566–1581, 2006.

[35] Wagner, A., Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis, In ECAI 2000 Workshop on Ontology Learning, 2000.

[36] Wei Li and McCallum, A., Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, In Proceedings of the International Conference on Machine Learning, 2006.

[37] Zavitsanos, E., Paliouras, G., Vouros, G.A., and Petridis, S., Discovering Subsumption Hierarchies of Ontology Concepts from Text Corpora, IEEE/WIC/ACM International Conference on Web Intelligence – WI '07, 2007.