

Determining Automatically the Size of Learned Ontologies

Elias Zavitsanos^{1,2} and Sergios Petridis¹ and Georgios Paliouras¹ and George A. Vouros²

Abstract. Determining the size of an ontology that is automatically learned from texts is an open issue. In this paper, we study the similarity between ontology concepts at different levels of a taxonomy, quantifying in a natural manner the quality of the ontology attained. Our approach is integrated in a method for language-neutral learning of ontologies from texts, which relies on conditional independence tests over thematic topics that are discovered using LDA.

1 INTRODUCTION

Ontology learning is commonly viewed [1, 3] as the task of *extending* or *enriching* a seed ontology with new ontology elements mined from text corpora. While much work concentrates on enriching existing ontologies, in this paper, we propose an automated statistical approach to ontology learning, without presupposing the existence of a seed ontology. The proposed method tackles both tasks of concept identification and taxonomy construction. Among the difficulties of such an endeavor, is the determination of the appropriate depth of the subsumption hierarchy, given the text collection at hand. The benefit of being able to determine the depth of a taxonomy is that the hierarchy captures accurately the domain knowledge provided by the texts, reducing the extent of overlap among concepts and providing a coherent representation of the domain.

In the proposed method, concepts are identified and represented as multinomial distributions over terms in documents, using the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [4], following the Latent Dirichlet Allocation (LDA) [2] model. To discover the subsumption relations between the identified concepts, conditional independence tests among these concepts are performed. Finally, statistical measures between the discovered concepts at different levels of the hierarchy are used to optimize the size of the ontology.

2 THE PROPOSED METHOD

Given a corpus of documents, treating each document as a bag of words, we remove the stop-words. The remaining words form the term space for the application of the topic generation model (LDA). The next step creates a Document - Term matrix, each entry of which records the frequency of each term in each document. This matrix is used as input to LDA.

Next, the iterative task of the learning method is initiated. Sets of topics, that we call layers, are generated by the iterative application of LDA. Starting with one topic and by incrementing the number

of topics in each iteration, layers with more topics are generated. A layer comprising few topics attempts to capture all the knowledge of the corpus through generic topics. As the number of topics increases, the topics become more focused, capturing more detailed domain knowledge. Thus, the method starts from “general” topics, iterates, and converges to more “specific” ones.

In each iteration, the method identifies the subsumption relations that hold between topics of different layers according to their conditional independencies. Since the generated topics are random variables, e.g. A and B , by measuring their mutual information we obtain an estimate of their mutual dependence. Given a third variable C that makes A and B conditionally independent, the mutual information of topics A and B is reduced and is captured by topic C , i.e., C is a broader topic than the others. Thus, we may safely assume that C subsumes both A and B and the corresponding relations are added to the ontology. Moreover, C has been generated before A and B . Thus, it belongs in a layer that contains topics that are broader in meaning than the ones in the layer of A and B .

A significant contribution is the determination of the appropriate depth of the hierarchy from the given corpus of documents. We use a criterion based on the similarity of topic distributions that indicates the convergence towards the appropriate depth. We thus improve on our recent work [5] by fitting this criterion, which controls the iterative process of the topic discovery.

This stopping criterion is based on the symmetric KL divergence between concepts of different levels that participate in subsumption relations. The intuition is that the KL divergence between concepts that belong in the top levels of the hierarchy should be higher than the KL divergence between concepts that belong in the lower levels. This is because the top concepts are broader in scope than lower ones and the “semantic distance” between them and their children is expected to be higher than this of more specific concepts and their children.

To validate this assumption, we have experimented with the Genia³ and the Lonely Planet gold ontologies and the corresponding corpora⁴. In order to measure the similarity of the concepts in the ontologies using statistical measures, we represented the concepts of each gold standard ontology as probability distributions over the term space of the corresponding corpus. To create such a representation, we have to measure the frequency of the terms that appear in the context of each concept. In both corpora, the concept instances are annotated in the texts, providing direct population of the concepts in the golden standard ontologies with their instances. Therefore, it is possible to associate each document to the concept(s) that it refers to, by counting the concept instances that appear in the document.

¹ Inst. of Informatics and Telecommunications, NCSR “Demokritos”, Greece, email: {izavits | petridis | paliourg}@iit.demokritos.gr

² University of Aegean, Dpt. of Information and Communication Systems Engineering, Greece, email: georgev@aegean.gr

³ The GENIA project, <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA>

⁴ The Lonely Planet travel advise and information, <http://www.lonelyplanet.com/>

Thus, we create feature vectors based on the document in which each concept appears. These feature vectors form a two-dimensional matrix that records the frequency of each term in the context of each concept. That is, we have a representation of each concept as a distribution over the term space of the text collection. For each concept, frequencies are normalized giving a probability distribution over the term space.

Figure 1 depicts the results obtained by measuring the similarity between concepts that participate in subsumption relations, in the case of the Genia and the Lonely Planet gold standard ontologies. Small values of KL divergence indicate high similarity between concepts. Figure 1 also confirms our assumption that concepts at the lower levels of the hierarchy are more similar to their children than concepts at higher levels of the hierarchy.

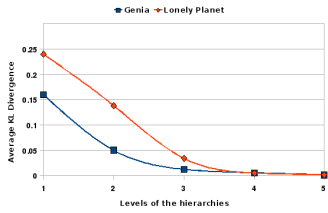


Figure 1. Average KL Divergence of subsumed concepts in the Genia and the Lonely Planet gold standard ontology.

Based on this approach, we define a relative criterion that indicates how deep the hierarchy should be according to the information provided by the corpus of documents. This criterion, which controls the iterative task of the proposed method is defined as:

$$1 - \frac{KL_{bottom}}{KL_{top}} < \varepsilon. KL_{top}$$

corresponds to the average symmetric KL divergence between the concepts of level l and the concepts of level $l+1$. KL_{bottom} is the average symmetric KL divergence between the concepts at level $l+1$ and the concepts of level $l+2$. Values close to 0 indicate that the new level of concepts added does not differ much from the parent concepts. Thus we are reaching maximum “specificity” and therefore optimal depth. Actually, the parameter ε has a very small value very close to zero to avoid small rounding errors during the computations.

3 EVALUATION

We have evaluated the proposed method on both corpora introduced in section 2. Our evaluation procedure uses the representation of the golden standard concepts as probability distributions over the term space of the documents, as explained in section 2. In addition, the concepts of the produced hierarchy have exactly the same representation. They are probability distributions over the same term space. We can, thus, perform a one-to-one comparison of the golden concepts and the produced topics. Specifically, a topic is matched to a concept if their corresponding distributions were the “closest” compared to all the other and their KL divergence was below a fixed threshold th_{KL} .

The quantitative results have been produced using the metrics of *Precision* and *Recall*. The choice of threshold th_{KL} affects the quantitative results, since a strict choice would force few topics to be matched with golden concepts, while a loose choice would cause many topics to be matched with golden concepts. We have chosen a value of $th_{KL} = 0.2$ for the purposes of our evaluation, as we

observed relative insensitivity of the result for values between 0.2 and 0.4 and we opted for the more conservative value in this plateau. Table 1 depicts the results.

Table 1. Evaluation results for the Genia and the Lonely Planet corpora.

Concept Identification for the Genia corpus		
Precision	Recall	F-measure
94%	76%	84%
Subsumption Hierarchy Construction for the Genia corpus		
Precision	Recall	F-measure
93%	75%	83%
Concept Identification for the Lonely Planet corpus		
Precision	Recall	F-measure
62%	36%	44%
Subsumption Hierarchy Construction for the Lonely Planet corpus		
Precision	Recall	F-measure
53%	35%	42%

To obtain a more detailed picture of the performance of the method, we replaced the stopping criterion with predefined depths for the learned hierarchy and we experimented in both corpora. Figure 2 presents the evaluation results in terms of the F-measure for various depths of the hierarchy, using the same evaluation style.

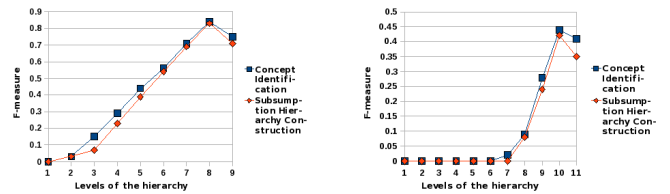


Figure 2. F-measures for Concepts Identification and Subsumption Hierarchy Construction for the Genia (left) and the Lonely Planet (right) corpora.

Figure 2 depicts that for a predefined depth of 8 levels in case of Genia, or 10 levels, in the case of Lonely Planet, the F-measure is maximized reaching the values of table 1. Therefore, the method determined correctly the appropriate depth in both corpora.

ACKNOWLEDGEMENTS

The presented work was supported by the research and development project ONTOSUM⁵, funded by the Greek General Secretariat for Research and Technology.

REFERENCES

- [1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, ‘Enriching very large ontologies using the www’, in *Ontology Construction Workshop*, (2000).
- [2] D.M. Blei, A.Y. Ng, and M.I. Jordan, ‘Latent dirichlet allocation’, *Journal of Machine Learning Research*, (2003).
- [3] A. Faatz and R. Steinmetz, ‘Ontology enrichment with texts from the www’, in *Semantic Web Mining Workshop ECML/PKDD*, (2002).
- [4] T. Griffiths and M. Steyvers, ‘A probabilistic approach to semantic representation’, in *Conference of the Cognitive Science Society*, (2002).
- [5] E. Zavitsanos, G. Paliouras, G.A. Vouros, and S. Petridis, ‘Discovering subsumption hierarchies of ontology concepts from text corpora’, in *Proceedings of the International Conference on Web Intelligence*, (2007).

⁵ See also <http://www.ontosum.org/>