Chapter 10
Exploiting Transitivity in Probabilistic Models for Ontology Learning

Francesca Fallucchi
University of Rome – Guglielmo Marconi, Italy

Fabio Massimo Zanzotto
University of Rome – Tor Vergata, Italy

ABSTRACT

Capturing word meaning is one of the challenges of Natural Language Processing (NLP). Formal models of meaning such as ontologies are knowledge repositories used in a variety of applications. To be effectively used, these ontologies have to be large or, at least, adapted to specific domains. This chapter's main goal is to practically contribute to the research on ontology learning models by covering different aspects of the task.

The authors propose probabilistic models for learning ontologies that expand existing ontologies taking into account both corpus-extracted evidence and the structure of the generated ontologies. The model exploits structural properties of target relations such as transitivity during learning. They then propose two extensions of the probabilistic models: a model for learning from a generic domain that can be exploited to extract new information in a specific domain and an incremental ontology learning system that puts human validations in the learning loop. This latter provides a graphical user interface and a human-computer interaction workflow supporting the incremental leaning loop.

INTRODUCTION

Gottfried Wilhelm Leibniz was convinced that human knowledge was like a *"bazaar"*: a place full of all sorts of goods without any order or inventory. As in a *"bazaar*," searching a little piece of specific knowledge is a challenge that can last forever. Nowadays, we have powerful machines to process and collect data. These machines, combined with the human need of exchanging and sharing information, produced an incredibly large evolving collection of documents, partially shared with the World Wide Web. The Web is a

DOI: 10.4018/978-1-4666-0188-8.ch010

modern worldwide scale knowledge *"bazaar"* full of any sort of information where searching specific information is a titanic task.

Ontologies represent the Semantic Web's reply to the need of searching knowledge in the Web. These ontologies provide shared metadata vocabularies (Berners-Lee, Hendler, & Lassila, 2001). Data, documents, images, and information sources in general, described through these vocabularies, will be thus accessible as organized with explicit semantic references for humans as well as for machines. Yet, to be useful, ontologies should cover large part of human knowledge. Automatically learning these ontologies from document collections is the major challenge.

Models for automatically learning semantic networks of words from texts use both corpus-extracted evidences and existing language resources (Basili, Gliozzo, & Pennacchiotti, 2007). All these models rely on two hypotheses: *Distributional Hypothesis* (*DH*) (Harris, 1964) and *Lexico-Syntactic Patterns exploitation hypothesis* (LSP) (Robison, 1970). While these are powerful tools to extract relations among concepts using texts, models based on these hypotheses do not explicitly exploit structural properties of target relations when learning taxonomies or semantic networks of words. DH models intrinsically use structural properties of semantic networks of words such as transitivity, but these models cannot be applied for learning transitive semantic relations other than the generalization. LSP models are interesting because they can learn any kind of semantic relations. Yet, these models do not exploit structural properties of target relations when learning taxonomies or semantic networks of words. In general, structural properties of semantic networks of words, when relevant, are not used in machine learning models to better induce confidence values for extracted semantic relations. Even where transitivity is explicitly used (Snow, Jurafsky, & Ng, 2006), it is not directly exploited to model confidence values. It is only used in an iterative maximization process of the probability of the entire semantic network. In this chapter, we propose a probabilistic approach that exploits LSP hypothesis and formally includes the exploitation of transitivity during learning.

Probabilistic models for learning semantic networks exploiting transitivity do not completely solve the problem of learning semantic networks. We have a second problem to tackle. When dealing with learning semantic networks of words from texts such as learning ontologies, we generally have *ontology-rich* domains with large structured domain knowledge repositories or large general corpora with large general structured knowledge repositories such as WordNet (Miller, 1995). Systems that automatically create, adapt, or extend existing semantic networks of words need a sufficiently large number of documents and existing structured knowledge to achieve reasonable performance. Thus, it is generally possible to extract good probabilistic models for *ontology-rich* domains or the general language. When building semantic networks for *ontology-poor* domains, we then need to rely on probabilistic models learnt out-of-domain or for the general language. If the target domain has not relevant pre-existing semantic networks of words to expand, we will not have enough data for training the initial model. In general, in learning methods the amount of out-of-domain data is larger than in-domain data. For this reason, in this chapter we present methods that, with a small effort for the adaptation to different specific knowledge domains, can exploit out-of-domain data for building in-domain models with bigger accuracy.

Finally, when learning semantic networks, we need to put human validations in the loop. Systems for creating or augmenting semantic networks of words using information extracted from texts need a manual validation for assessing the quality of semantic networks of words expansion. Yet, these systems do not use the manual validation for refining the information extraction model that proposes novel links in the networks. Manual validation can be efficiently exploited if used in