Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015) 3–21

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Global machine learning for spatial ontology population

Parisa Kordjamshidi *, Marie-Francine Moens

KU Leuven, Belgium

A R T I C L E I N F O

Article history:
Received 15 July 2013
Received in revised form
21 May 2014
Accepted 18 June 2014
Available online 26 June 2014

Keywords:
Spatial information extraction
Spatial role labeling
Text mining
Structured output learning
Ontology population
Natural language processing

A B S T R A C T

Understanding spatial language is important in many applications such as geographical information systems, human computer interaction or text-to-scene conversion. Due to the challenges of designing spatial ontologies, the extraction of spatial information from natural language still has to be placed in a well-defined framework. In this work, we propose an ontology which bridges between cognitive–linguistic spatial concepts in natural language and multiple qualitative spatial representation and reasoning models. To make a mapping between natural language and the spatial ontology, we propose a novel global machine learning framework for ontology population. In this framework we consider relational features and background knowledge which originate from both ontological relationships between the concepts and the structure of the spatial language. The advantage of the proposed global learning model is the scalability of the inference, and the flexibility for automatically describing text with arbitrary semantic labels that form a structured ontological representation of its content. The machine learning framework is evaluated with SemEval-2012 and SemEval-2013 data from the spatial role labeling task.

1. Introduction

An essential function of natural language is to talk about the location and translocation of objects in space. Understanding spatial language is important in many applications such as geographical information systems (GIS), human computer interaction, text-to-scene conversion, and representation and extraction of spatial information from web resources such as travelers blogs or websites about tourism. Due to the complexity of spatial primitives and notions, and the challenges of designing ontologies for formal spatial representation, the extraction of the spatial semantics from natural language still has to be placed in a well-defined framework.

We have two main contributions toward solving this problem. The first contribution is that we propose a spatial ontology based on two layers of semantics. This ontology is based on a previously proposed spatial annotation scheme by the authors [1]. Its first layer is based on commonly accepted cognitive spatial notions and the second is based on multiple well-known qualitative spatial reasoning models. An automatic mapping to such an ontology bridges between natural language and qualitative spatial representation and reasoning models, which makes automatic spatial reasoning based on spatial information in linguistic expressions feasible. This ontology can be integrated in larger ontologies, for example, to represent spatial meaning in unstructured data in the context of the Semantic Web.

The second contribution of this work is that we propose a novel global supervised machine learning model for spatial ontology population. For this supervised learning framework, we build rich annotated corpora and an evaluation scheme. We point to the linguistic features and structural characteristics of spatial language that aid the use of machine learning. We view ontology population as a means for creating meaning representations from text. In this model the segments of the input text are described by semantic abstractions or concepts and their relationships defined by the ontology, which form the output space of the learning problem [2]. In the proposed global learning framework, the ontology components including spatial roles and their relations, and multiple formal semantic types are learned while taking into account the ontological constraints and the structural characteristics of the spatial language.

Learning a model that considers the global correlations between the output components usually becomes computationally complex. To deal with the complexity in training and prediction phases, we use an efficient inference approach based upon combinatorial optimization techniques for both phases. This approach can deal with a large number of variables and constraints, and makes building a structured machine learning model for ontology population, feasible.

We decompose the learning problem into simpler problems that are jointly optimized. We propose a technique which we call

* Corresponding author. Tel.: +32 486517941.
E-mail address: parisa.kordjamshidi@cs.kuleuven.be (P. Kordjamshidi).

http://dx.doi.org/10.1016/j.websem.2014.06.001
1570-8268/

communicative inference based on the ideas of *alternating optimization* for solving smaller subproblems of the main objective function [3]. Each subproblem is solved by using linear programming (LP) solvers and the subproblems communicate to each other by passing the local solutions. We show that the suggested framework is beneficial compared to local learning as well as compared to pipelining the independently learned models for the concepts in the ontology. The proposed inference approach makes the global learning scalable.

The application of the global machine learning model for ontology population is not limited to the extraction of spatial semantics; it could be used to populate any ontology. Moreover, due to decomposing the ontology to its solvable parts, this approach is scalable to be applied for approximate global learning for large ontologies of the Semantic Web. We argue therefore that this work is an important step towards automatically describing text with semantic labels that form a structured ontological representation of the content.

Our extensive experimental study using the spatial ontology indicates the advantage of global learning while considering ontological constraints and structural characteristics of the spatial language compared to learning local models for the various parts of the ontology independently. The experiments are performed using the corpora provided by the SemEval-2012 and SemEval-2013 shared task on spatial role labeling.

In Section 2, we provide the problem definition and the spatial ontology population task in its two layers of semantics. In Section 3, we discuss the features and constraints that are useful for learning the spatial ontology population. A background to structured learning is provided in Section 4. The proposed structured learning model for spatial ontology population is described in Section 5. The proposed inference approach is explained in Section 6. Section 7 specifies the details of the components of the spatial ontology population model. The various designed local and global models are clarified in Section 8. Section 9 presents the experimental results. An overview of the related research is provided in Section 10. We draw conclusions, set our work in a broader context, and point to the future extensions in Section 11.

2. General problem definition

We define a framework for mapping natural language to spatial ontologies. Although pragmatic, our proposed framework is based on the theoretical cognitive and linguistic foundations, as well as on cognitively adequate formal spatial models. The task is formulated as an ontology population to be performed via supervised machine learning models. We aim at learning to assign the segments in the sentence to the concepts in the ontology. The considered concepts form a *light weight* ontology which is based on a previously proposed spatial annotation scheme by the authors [1]. We highlight the distinction between two *spatial role labeling* (SpRL) and *spatial qualitative labeling* (SpQL) layers in the ontology. We describe the structural characteristics of the two-layered ontology to be exploited in the learning models.

2.1. Two layers of semantics

Spatial language can convey complex spatial relations along with polysemy and ambiguity inherited in natural language. Linguistic spatial expressions can express various aspects of the space at the same time [4]. In contrast to natural language, formal spatial models focus on one particular spatial aspect such as orientation, topology or distance and specify its underlying spatial logic in detail [5,6]. Therefore there is a gap between the level of expressivity and specification of natural language and spatial calculi models [7].

Due to this gap, learning how to map the spatial information in natural language onto a formal representation is a challenging problem. However, such a mapping is useful because formal spatial models enable automatic spatial reasoning that is difficult to perform on natural language expressions. To overcome the complexity of this problem in a systematic way, our spatial ontology is divided into two abstraction layers [7–9]:

1. A layer of *linguistic conceptual representation* called spatial role labeling (SpRL), which predicts the existence of spatial information at the sentence level by identifying the words that play a particular spatial role as well as their spatial relationship [10];
2. A layer of *formal semantic representation* called spatial qualitative labeling (SpQL), in which the spatial relation is described with semantic attributes based on qualitative spatial representation models (QSR) [11,12].

In our conceptual model we argue that mapping the language to multiple spatial representation models could solve the problem of the existing gap between the two layers to some extent (also see [13,14] in the context of robotics and navigational instructions). Because various formal representations capture the semantics from different angles, their combination covers various aspects of spatial semantics needed for locating the objects in the physical space. Hence, the SpQL has to contain multiple calculi models with a practically acceptable level of generality. However, we believe that this two layered model does not yet yield sufficient flexibility for ideal spatial language understanding. As in any other semantic task in natural language, additional layers of *discourse* and *pragmatics* must be worked out, which is not the focus of this work.

2.2. Task definition as ontology population

Our main task is to map a given sentence x composed of a number of words $x_1 \dots x_n$ to the predefined spatial ontology shown in Fig. 1(a). The task is to label the words in the sentence with spatial roles (SpRL), detect the spatial relations, and label the spatial relations with their spatial semantics including coarse-grained as well as fine-grained semantic labels. The words can have multiple roles and the relations can have multiple semantic assignments. The labels are assigned according to the relationships and constraints that we discuss in the following sections. The considered spatial ontology here is only a lightweight [15] ontology, but pinpoints the main challenges in the recognition of ontological label structures in text.

2.2.1. Spatial role labeling (SpRL)

In the *spatial role labeling* (SpRL) layer the cognitive–linguistic spatial semantics based on the theory of *holistic spatial semantics* are considered [16,17]. Fig. 1(b) shows the sentence, *The flag of Paraguay is waving at the top of the building.*, which is labeled according to the nodes in the spatial ontology of Fig. 1(a). In the SpRL step the goal is to identify the words that play a spatial role in the sentence and to classify their roles; moreover to recognize the link between the spatial roles and extract the spatial relations. In this sentence, we need to extract a spatial relationship signaled by *at* that holds between *flag* and *building*. The word *flag* has the role of *trajector* (*tr*). The trajector is an entity whose location is described. The word *building* has the role of *landmark* (*lm*). The landmark is a reference object for describing the location of a trajector. These two spatial entities are related by the spatial expression *at* that is the *spatial indicator* (*sp*). The spatial indicator signals the existence of spatial information in the sentence.

These spatial roles are the three main nodes in our ontology. We refer to these nodes as *single label*s. A *single label* refers to an independent concept in the ontology. For a spatial configuration, we consider the link between the three roles, which is labeled as