Data & Knowledge Engineering 100 (2015) 211–225

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Approximate and selective reasoning on knowledge graphs: A distributional semantics approach

André Freitas [a,*], João C.P. da Silva [b,c], Edward Curry [b], Paul Buitelaar [b,d]

[a] Department of Computer Science and Mathematics, University of Passau, Dr-Hans-Kapfinger Strasse, 12, R.109, Passau, 94032, Germany
[b] Insight Centre for Data Analytics, National University of Ireland, Insight, IDA Buisiness Park, Lower Dangan, Galway, Ireland
[c] Computer Science Department, Federal University of Rio de Janeiro, Caixa-Postal 68530, CEP: 21941-916, Rio de Janeiro, RJ, Brazil
[d] College of Graduate Studies, University of South Africa, Pretoria, South Africa

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 24 June 2015
Accepted 25 June 2015
Available online 26 July 2015

Keywords:
Commonsense reasoning
Selective reasoning
Distributional semantics
Hybrid distributional-relation models
Semantic approximation

ABSTRACT

Tasks such as question answering and semantic search are dependent on the ability of querying and reasoning over large-scale commonsense knowledge bases (KBs). However, dealing with commonsense data demands coping with problems such as the increase in schema complexity, semantic inconsistency, incompleteness and scalability. This paper proposes a selective graph navigation mechanism based on a distributional relational semantic model which can be applied to querying and reasoning over heterogeneous knowledge bases (KBs). The approach can be used for approximative reasoning, querying and associational knowledge discovery. In this paper we focus on commonsense reasoning as the main motivational scenario for the approach. The approach focuses on addressing the following problems: (i) providing a semantic selection mechanism for facts which are relevant and meaningful in a specific reasoning and querying context and (ii) allowing coping with information incompleteness in large KBs. The approach is evaluated using ConceptNet as a commonsense KB, and achieved *high selectivity*, *high selectivity scalability* and *high accuracy in the selection of meaningful navigational paths*. Distributional semantics is also used as a principled mechanism to cope with information incompleteness.

1. Introduction

Building intelligent applications and addressing simple computational semantic tasks demand coping with large-scale commonsense knowledge bases (KBs). Querying and reasoning (Q&R) over large commonsense KBs are fundamental operations for tasks such as Question Answering, Semantic Search and Knowledge Discovery. However, in an open domain scenario, the scale of KBs and the number of direct and indirect associations between elements in the KB can make Q&R grow unmanageable. To the complexity of querying and reasoning over such large-scale KBs, it is possible to add the barriers involved in building KBs with the necessary consistency and completeness requirements.

With the evolution of open data, better information extraction frameworks and crowd-sourcing tools, large-scale structured KBs are becoming more available. This data can be used to provide commonsense knowledge for semantic applications. However, querying and reasoning over this data demands approaches which are able to cope with large-scale, semantically heterogeneous and incomplete KBs.

As a motivational scenario, suppose we have a KB with the following fact: *'John Smith is an Engineer'* and suppose the query *'Does John Smith have a degree?'* is issued over the KB. A complete KB would have the rule *'Every engineer has a degree'*, which would

* Corresponding author.
E-mail addresses: andre.freitas@uni-passau.de (A. Freitas), jcps.ufrj.br@gmail.com (J.C.P. da Silva).

http://dx.doi.org/10.1016/j.datak.2015.06.010
0169-023X/

materialize *'John Smith has a degree'*. For large-scale and open domain commonsense reasoning scenarios, model completeness and full materialization cannot be assumed. In this case, the information can be embedded in other facts in the KB (Fig. 1). The example sequence of relations between *engineer* and *degree* defines a path in a large-scale graph of relations between predicates, which is depicted in Fig. 1.

In a large-scale KB, full reasoning can become unfeasible. A commonsense KB would contain vast amounts of facts and a complete inference over the entire KB would not scale to its size. Furthermore, while the example path is a meaningful sequence of associations for answering the example query, there is a large number of paths which are not meaningful under a specific query context. In Fig. 1(1), for example, the reasoning path which goes through (1) is not related to the goal of the query (the relation between *engineer* and *degree*) and should be eliminated. Ideally a query and reasoning mechanism should be able to filter out facts and rules which are unrelated to the Q&R context. The ability to select the minimum set of facts which should be applied in order to answer a specific user information need is a fundamental element for enabling reasoning capabilities for large-scale commonsense knowledge bases.

Additionally, since information completeness of the KBs cannot be guaranteed, one missing fact in the KB would be sufficient to block the reasoning process. In Fig. 1(2) the lack of a fact connecting university and college eliminates the possibility of answering the query. Ideally Q&R mechanisms should be able to cope with some level of KB incompleteness, approximating and filling the gaps in the KBs.

This work proposes a *selective reasoning approach* which uses a *hybrid distributional-relational semantic model* to address the problems previously described. Distributional semantic models (DSMs) use statistical co-occurrence patterns, automatically extracted from large unstructured text corpora, to support the creation of comprehensive quantitative semantic models. In this work, DSMs are used as complementary semantic layers to the relational/logical model, which supports coping with semantic approximation and incompleteness. The proposed approach focuses on the following contributions:

- provision of a selective Q&R approach using a distributional semantics heuristics, which reduces the search space for large-scale KBs at the same time it maximizes paths which are more meaningful for a given reasoning context;
- definition of a Q&R model which copes with the information incompleteness present at the KB, using the distributional model to support semantic approximations, which can fill the lack of information in the KB during the reasoning process.

This work is organized as follows: Section 2 introduces natural language commonsense knowledge bases and briefly describes ConceptNet [11]; Section 3 provides an introduction on distributional semantics; Section 4 describes the τ-Space distributional-relational semantic model which is used for the selection reasoning mechanism; Section 5 defines the distributional representation for the commonsense *KB*; Section 6 describes the selective reasoning mechanism (*distributional navigational algorithm*); Section 7 provides an evaluation of the approach using explicit semantic analysis (ESA) as a distributional semantic model and ConceptNet [11] as KB; Section 8 describes related work and finally, Section 9 provides conclusions and future work.

2. Natural language commonsense knowledge bases (NLCS-KB)

More recent commonsense KBs such as ConceptNet are shifting from a logic representation framework to a natural language-based commonsense knowledge representation [11,14]. The motivation behind this change in perspective is to improve the scale

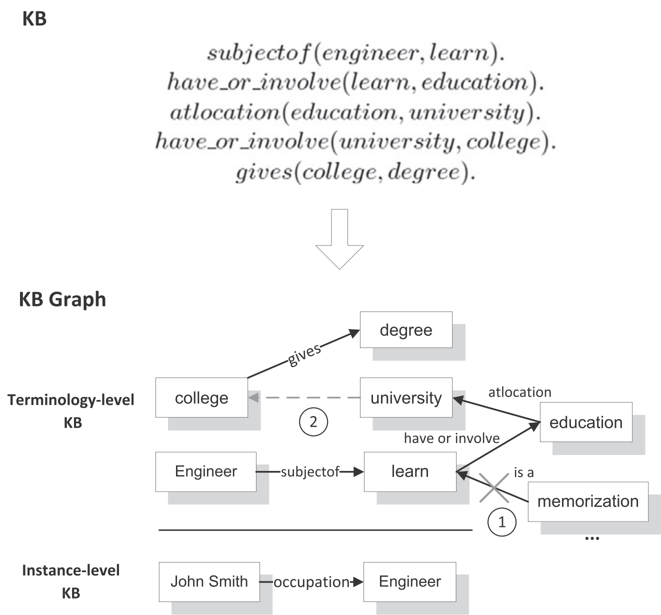

Fig. 1. (1) Selection of meaningful paths. (2) Coping with information incompleteness.