Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 112–131

Contents lists available at ScienceDirect

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

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

Sar-graphs: A language resource connecting linguistic knowledge with semantic relations from knowledge graphs

Sebastian Krause [a,*], Leonhard Hennig [a], Andrea Moro [b], Dirk Weissenborn [a], Feiyu Xu [a], Hans Uszkoreit [a], Roberto Navigli [b]

[a] DFKI Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany

[b] Dipartimento di Informatica, Sapienza Università di Roma, Viale Regina Elena 295, 00161 Roma, Italy

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
27 January 2016
Accepted 6 March 2016
Available online 15 March 2016

Keywords:
Knowledge graphs
Language resources
Linguistic patterns
Relation extraction

ABSTRACT

Recent years have seen a significant growth and increased usage of large-scale knowledge resources in both academic research and industry. We can distinguish two main types of knowledge resources: those that store factual information about entities in the form of semantic relations (e.g., Freebase), namely so-called knowledge graphs, and those that represent general linguistic knowledge (e.g., WordNet or UWN). In this article, we present a third type of knowledge resource which completes the picture by connecting the two first types. Instances of this resource are *graphs of semantically-associated relations (sar-graphs)*, whose purpose is to link semantic relations from factual knowledge graphs with their linguistic representations in human language.

We present a general method for constructing sar-graphs using a language- and relation-independent, distantly supervised approach which, apart from generic language processing tools, relies solely on the availability of a lexical semantic resource, providing sense information for words, as well as a knowledge base containing seed relation instances. Using these seeds, our method extracts, validates and merges relation-specific linguistic patterns from text to create sar-graphs. To cope with the noisily labeled data arising in a distantly supervised setting, we propose several automatic pattern confidence estimation strategies, and also show how manual supervision can be used to improve the quality of sar-graph instances. We demonstrate the applicability of our method by constructing sar-graphs for 25 semantic relations, of which we make a subset publicly available at http://sargraph.dfki.de.

We believe sar-graphs will prove to be useful linguistic resources for a wide variety of natural language processing tasks, and in particular for information extraction and knowledge base population. We illustrate their usefulness with experiments in relation extraction and in computer assisted language learning.

1. Introduction

Knowledge graphs are vast networks which store entities and their semantic types, properties and relations. In recent years considerable effort has been invested into constructing these large knowledge bases in academic research, community-driven projects and industrial development. Prominent examples include Freebase [1], Yago [2,3], DBpedia [4], NELL [5,6], WikiData [7], PROSPERA [8], Google's Knowledge Graph [9] and also the Google Knowledge Vault [10]. A parallel and in part independent development is the emergence of several large-scale knowledge resources with a more language-centered focus, such as UWN [11], BabelNet [12], ConceptNet [13], and UBY [14]. These resources are important contributions to the linked data movement, where repositories of world-knowledge and linguistic knowledge complement each other. In this article, we present a method that aims to bridge these two types of resources by automatically building an intermediate resource.

In comparison to (world-)knowledge graphs, the underlying representation and semantic models of linguistic knowledge resources exhibit a greater degree of diversity. ConceptNet makes use of natural-language representations for modeling common-sense information. BabelNet integrates entity information from Wikipedia with word senses from WordNet, as well as with many other resources such as Wikidata and Wiktionary [15].

* Corresponding author.
E-mail address: skrause@dfki.de (S. Krause).

http://dx.doi.org/10.1016/j.websem.2016.03.004
1570-8268/

UWN automatically builds a multilingual WordNet from various resources, similar to UBY, which integrates multiple resources via linking on the word-sense level. Few to none of the existing linguistic resources, however, provide a feasible approach to explicitly linking semantic relations from knowledge graphs with their linguistic representations. We aim to fill this gap with the resource whose structure we define in Section 2 and whose construction method we detail in Section 3. Instances of this resource are *graphs of semantically-associated relations*, which we refer to by the name *sar-graphs*. Our definition is a formalization of the idea sketched in [16]. We believe that sar-graphs are examples for a new type of knowledge repository, *language graphs*, as they represent the linguistic patterns for relations in a knowledge graph. A language graph can be thought of as a bridge between the language and knowledge encoded in a knowledge graph, a bridge that characterizes the ways in which a language can express instances of one or several relations, and thus a mapping between strings and things.

The construction strategies of the described (world-)knowledge resources range from (1) integrating existing structured or semi-structured knowledge (e.g., Wikipedia infoboxes) via (2) crowd-sourcing to (3) automatic extraction from semi- and unstructured resources, where often (4) combinations of these are implemented. At the same time the existence of knowledge graphs enabled the development of new technologies for knowledge engineering, e.g., distantly supervised machine-learning methods [8,17–20]. Relation extraction is one of the central technologies contributing to the automatic creation of fact databases [10], on the other hand it benefits from the growing number of available factual resources by using them for automatic training and improvement of extraction systems. In Section 3, we describe how our own existing methods [18], which exploit factual knowledge bases for the automatic gathering of linguistic constructions, can be employed for the purpose of sar-graphs. Then in turn, one of many potential applications of sar-graphs is relation extraction, which we illustrate in Section 7.

An important aspect of the construction of sar-graphs is the disambiguation of their content words with respect to lexical semantics knowledge repositories, thereby generalizing content words with word senses. In addition to making sar-graphs more adjustable to the varying granularity needs of possible applications, this positions sar-graphs as a link hub between a number of formerly independent resources (see Fig. 1). Sar-graphs represent linguistic constructions for semantic relations from factual knowledge bases and incorporate linguistic structures extracted from mentions of knowledge-graph facts in free texts, while at the same time anchoring this information in lexical semantic resources. We go into further detail on this matter in Section 6.

The distantly supervised nature of the proposed construction methodology requires means for automatic and manual confidence estimation for the extracted linguistic structures, presented in Section 4. This is of particular importance when unstructured web texts are exploited for finding linguistic patterns which express semantic relations. Our contribution is the combination of battle-tested confidence-estimation strategies [18,21] with a large manual verification effort for linguistic structures. In our experiments (Section 5), we continue from our earlier work [18,22], i.e., we employ Freebase as our source of semantic relations and the lexical knowledge base BabelNet for linking word senses. We create sar-graphs for 25 relations, which exemplifies the feasibility of the proposed method, also we make the resource publicly available for this core set of relations.

We demonstrate the usefulness of sar-graphs by applying them to the task of relation extraction, where we identify and compose mentions of argument entities and projections of n-ary semantic relations. We believe that sar-graphs will prove to be a valuable resource for numerous other applications, such as adaptation of parsers to special recognition tasks, text summarization, language generation, query analysis and even interpretation of telegraphic style in highly elliptical texts as found in SMS, Twitter, headlines or brief spoken queries. We therefore make this resource freely available to the community, and hope that other parties will find it of interest (Section 8).

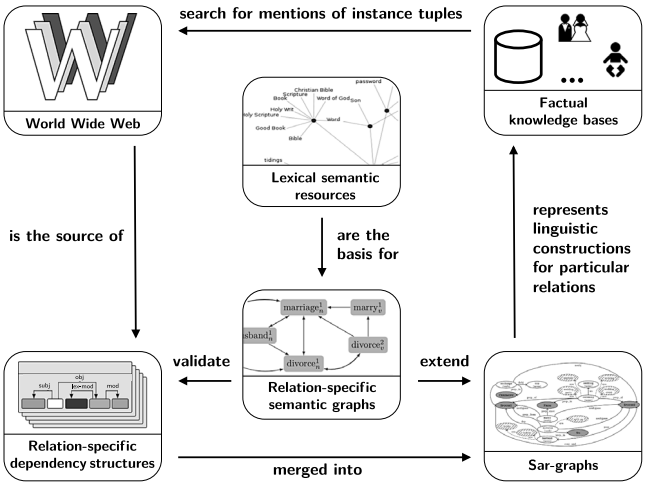

Fig. 1. Relation of sar-graphs to other knowledge resources.

2. Sar-graphs: a linguistic knowledge resource

Sar-graphs [16] extend the current range of knowledge graphs, which represent factual, relational and common-sense information for one or more languages, with linguistic knowledge, namely, linguistic variants of how semantic relations between abstract concepts and real-world entities are expressed in natural language text.

2.1. Definition

Sar-graphs are directed multigraphs containing linguistic knowledge at the syntactic and lexical semantic level. A sar-graph is a tuple

$$G_{r,l} = (V, E, s, t, f, A_f, \Sigma_f),$$

where

- V is the set of vertices,
- E is the set of edges,
- $s : E \mapsto V$ maps edges to their start vertex,
- $t : E \mapsto V$ maps edges to their target vertex.

As both vertices and edges are labeled, we also need an appropriate labeling function, denoted by f. f does more than just attaching atomic labels to edges and vertices but rather associates both with sets of features (i.e., attribute–value pairs) to account for the needed complexity of linguistic description:

$$f : V \cup E \mapsto \mathcal{P}(A_f \times \Sigma_f)$$

where

- $\mathcal{P}(\cdot)$ constructs a powerset,
- A_f is the set of attributes (i.e., attribute names) which vertices and edges may have, and
- Σ_f is the value alphabet of the features, i.e., the set of possible attribute values for all attributes.