Sar-graphs: A Linked Linguistic Knowledge Resource Connecting Facts with Language

Sebastian Krause, Leonhard Hennig, Aleksandra Gabryszak, Feiyu Xu, Hans Uszkoreit
DFKI Language Technology Lab, Berlin, Germany
{skrause,lehe02,alga02,feiyu,uszkoreit}@dfki.de

Abstract

We present *sar-graphs*, a knowledge resource that links semantic relations from factual knowledge graphs to the linguistic patterns with which a language can express instances of these relations. Sar-graphs expand upon existing lexico-semantic resources by modeling syntactic and semantic information at the level of relations, and are hence useful for tasks such as knowledge base population and relation extraction. We present a language-independent method to automatically construct sar-graph instances that is based on distantly supervised relation extraction. We link sar-graphs at the lexical level to BabelNet, WordNet and UBY, and present our ongoing work on pattern- and relation-level linking to FrameNet. An initial dataset of English sar-graphs for 25 relations is made publicly available, together with a Java-based API.

1 Introduction

Knowledge graphs, such as Freebase or YAGO, are networks which contain information about real-world 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 (Bollacker et al., 2008; Suchanek et al., 2008; Lehmann et al., 2015). A parallel and in part independent development is the emergence of large-scale lexical-semantic resources, such as BabelNet or UBY, which encode linguistic information about words and their relations (de Melo and Weikum, 2009; Navigli and Ponzetto, 2012; Gurevych et al., 2012). Both types of resources are important contributions to the linguistic linked open data movement, since they address complementary aspects of encyclopedic and linguistic knowledge.

Few to none of the existing resources, however, explicitly link the semantic relations of knowledge graphs to the linguistic patterns, at the level of phrases or sentences, that are used to express these relations in natural language text. Lexical-semantic resources focus on linkage at the level of individual lexical items. For example, BabelNet integrates entity information from Wikipedia with word senses from WordNet, UWN is a multilingual WordNet built from various resources, and UBY integrates several linguistic resources by linking them at the word-sense level. Linguistic knowledge resources that go beyond the level of lexical items are scarce and of limited coverage due to significant investment of human effort and expertise required for their construction. Among these are FrameNet (Baker et al., 1998), which provides fine-grained semantic relations of predicates and their arguments, and VerbNet (Schuler, 2005), which models verb-class specific syntactic and semantic preferences. What is missing, therefore, is a large-scale, preferably automatically constructed linguistic resource that links language expressions at the phrase or sentence level to the semantic relations of knowledge bases, as well as to existing terminological resources. Such a repository would be very useful for many information extraction tasks, e.g., for relation extraction and knowledge base population.

We aim to fill this gap with a resource whose structure we define in Section 2. Instances of this resource are *graphs of semantically-associated relations*, which we refer to by the name *sar-graphs*. We believe that sar-graphs are examples for a new type of knowledge repository, *language graphs*, as they represent the linguistic patterns for the relations contained in a knowledge graph. A language graph can be thought of as a bridge between

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 30–38,
Beijing, China, July 31, 2015.

the language and the facts encoded in a knowledge graph, a bridge that characterizes the ways in which a language can express instances of relations. Our contributions in this paper are as follows:

- We present a model for *sar-graphs*, a resource of linked linguistic patterns which are used to express factual information from knowledge graphs in natural language text. We model these patterns at a fine-grained lexico-syntactic and semantic level (Section 2).
- We describe the word-level linking of sar-graph patterns to existing lexical-semantic resources (BabelNet, WordNet, and UBY; Section 3)
- We discuss our ongoing work of linking sar-graphs at the pattern and relation level to FrameNet (Section 4)
- We describe a language-independent, distantly supervised approach for automatically constructing sar-graph instances, and present a first published and linked dataset of English sar-graphs for 25 Freebase relations (Section 5)

2 Sar-graphs: A linguistic knowledge resource

Sar-graphs (Uszkoreit and Xu, 2013) extend the current range of knowledge graphs, which represent factual, relational and common-sense information for one or more languages, with linguistic variants of how semantic relations between real-world entities are expressed in natural language.

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

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

where V is the set of vertices and E is the set of edges. The labeling function f associates both vertices and edges with sets of features (i.e., attribute-value pairs):

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

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.

The function of sar-graphs is to represent the linguistic constructions a language l provides for referring to instances of r. A vertex $v \in V$ corresponds to either a word in such a construction, or an argument of the relation. The features assigned to a vertex via the labeling function f provide information about lexico-syntactic aspects (*word form* and *lemma*, *word class*), and lexical semantics (*word sense*), or semantic attributes (*global entity identifier*, *entity type*, *semantic role in the target relation*). They may also provide statistical and meta information (e.g., *frequency*). The linguistic constructions are modeled as sub-trees of dependency-graph representations of sentences. We will refer to these trees as *dependency structures* or *dependency constructions*. Each structure typically describes one particular way to express relation r in language l. Edges $e \in E$ are consequently labeled with dependency tags, in addition to, e.g., frequency information.

A given graph instance is specific to a language l and target relation r. In general, r links $n \geq 2$ entities. An example relation is *marriage*, connecting two spouses to one another, and optionally to the location and date of their wedding, as well as to their date of divorce:

$$r_{mar.}(\textsc{Spouse1}, \textsc{Spouse2}, \textsc{Ceremony}, \textsc{From}, \textsc{To}).$$

If a given language l only provides a single construction to express an instance of r, then the dependency structure of this construction forms the entire sar-graph. But if the language offers alternatives to this construction, i.e., paraphrases, their dependency structures are also added to the sar-graph. They are connected in such a way that all vertices labeled by the same argument name are merged, i.e., lexical specifics like word form, lemma, class, etc. are dropped from the vertices corresponding to the semantic arguments of the target relation. The granularity of such a dependency-structure merge is however not fixed and can be adapted to application needs.

Figure 1 presents a sar-graph for five English constructions with mentions of the *marriage* relation. The graph covers the target relation relevant parts of the individual mentions, assembled stepwise in a bottom-up fashion. Consider the two sentences in the top-left corner of the figure:

Example 1

- *I met Eve's husband Jack.*
- *Lucy and Peter are married since 2011.*

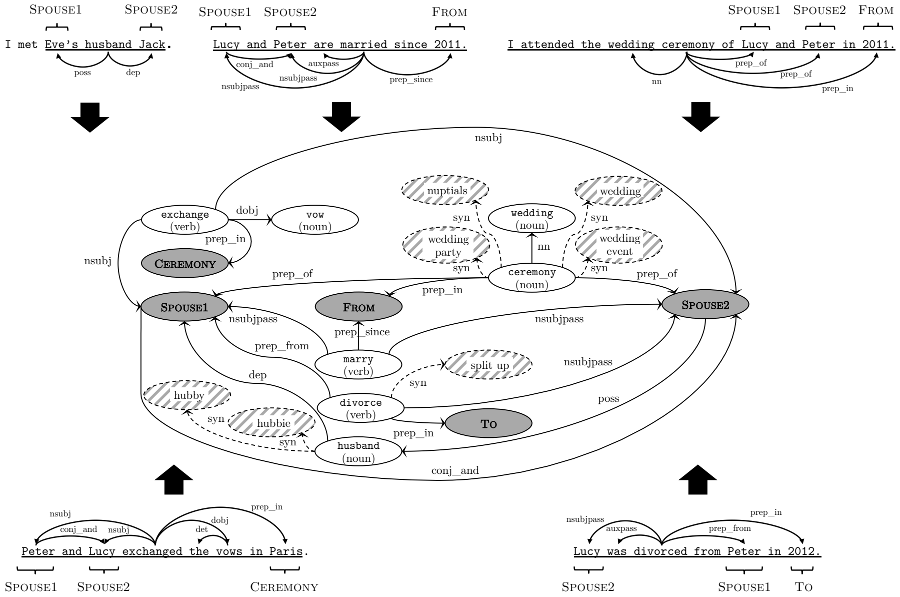

Figure 1: Example sar-graph for the *marriage* relation, constructed using the dependency patterns extracted from the sentences shown in the figure. Dashed vertices and edges represent additional graph elements obtained by linking lexical vertices to BabelNet.

From the dependency parse trees of these sentences, we can extract two graphs that connect the relation's arguments. The first sentence lists the spouses with a possessive construction, the second sentence using a conjunction. In addition, the second sentence provides the marriage date. The graph we extract from the latter sentence hence includes the dependency arcs *nsubjpass* and *prep_since*, as well as the node for the content word *marry*. We connect the two extracted structures by their shared semantic arguments, namely, SPOUSE1 and SPOUSE2. As a result, the graph in Figure 1 contains a path from SPOUSE1 to SPOUSE2 via the node *husband* for sentence (1), and an edge *conj_and* from SPOUSE1 to SPOUSE2 for sentence (2). The dependency relations connecting the FROM argument yield the remainder of the sar-graph.

The remaining three sentences from the figure provide alternative linguistic constructions, as well as the additional arguments CEREMONY and TO. The graph includes the paraphrases *exchange vows*, *wedding ceremony of*, and *was divorced from*. Note that both sentence (2) and (4) utilize a *conj_and* to connect the SPOUSES. The sar-graph includes this information as a single edge, but we can encode the frequency information as an edge attribute.

Less explicit relation mentions A key property of sar-graphs is that they store linguistic structures with varying degrees of explicitness wrt. to the underlying semantic relations. Constructions that refer to some part or aspect of the relation would normally be seen as sufficient evidence of an instance even if there could be contexts in which this implication is canceled:

Example 2

- *Joan and Edward exchanged rings in 2011.*
- *Joan and Edward exchanged rings during the rehearsal of the ceremony.*

Other constructions refer to relations that entail the target relations without being part of it:

Example 3

- *Joan and Edward celebrated their 12th wedding anniversary.*
- *Joan and Edward got divorced in 2011.*

3 Word-level linking

We link sar-graphs to existing linguistic linked open data (LOD) resources on the lexical level by mapping content word vertices to the lexical semantic resource BabelNet (Navigli and Ponzetto, 2012), and via BabelNet to WordNet and UBY-OmegaWiki. BabelNet is a large-scale multilingual semantic network automatically constructed from resources such as Wikipedia and WordNet. Its core components are Babel synsets, which are

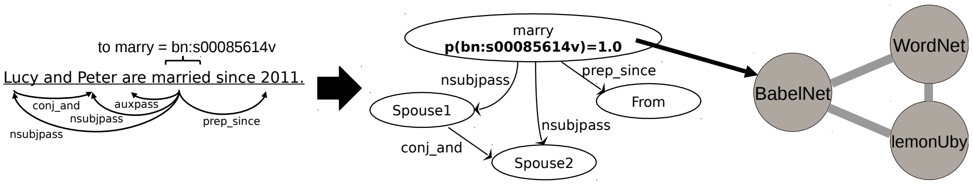

Figure 2: A minimal sar-graph disambiguation example, consisting of a single pattern, where the lexical vertex *marry* is disambiguated and linked to BabelNet, UBY, and WordNet.

sets of multilingual synonyms. Each Babel synset is related to other Babel synsets via semantic relations such as hypernymy, meronymy and semantic relatedness. BabelNet contains roughly 13M synsets, 117M lexicalizations and 354M relation instances.

Besides connecting sar-graphs to the linguistic LOD cloud, this mapping allows us to augment the lexico-syntactic and semantic information specified in sar-graphs with lexical semantic knowledge from the linked resources. In particular, we introduce new vertices for synonyms, and add new edges based on the lexical semantic relations specified in BabelNet. In Figure 1, these additional graph elements are represented as dashed vertices and edges.

To link sar-graph vertices to Babelnet, we disambiguate content words in our pattern extraction pipeline (see Section 5), using the graph-based approach described by Moro et al. (2014). The disambiguation is performed on a per-sentence basis, considering all content words in the sentence as potentially ambiguous mentions if they correspond to at least one candidate meaning in BabelNet. This includes multi-token sequences containing at least one noun. The candidate senses (synset identifiers) of all mentions in a sentence are linked to each other via their BabelNet relations to create a graph. The approach then iteratively prunes low-probability candidate senses from the graph to select the synset assignment that maximizes the semantic agreement within a given sentence. Once we have found this disambiguation assignment, we can use BabelNet's existing synset mappings to link each mention to its corresponding synsets in UBY-OmegaWiki and in the original Princeton WordNet. Figure 2 illustrates the word-level linking.

After extracting a dependency pattern from a given sentence, we store the synset assignments as a property for each content word vertex of the pattern. In the final, merged sar-graph, each content word vertex is hence associated with a distribution over synset assignments, since the same pattern may occur in multiple source sentences, with potentially different local disambiguation decisions.

4 Alignment to FrameNet

In addition to the straightforward sense-level linking of sar-graphs to thesauri, we aim to establish connections at more abstract information layers, e.g., to valency lexicons. In this section, we present our ongoing efforts for aligning sar-graphs with FrameNet at the level of phrases and relations.

FrameNet The Berkeley FrameNet Project (Baker et al., 1998; Ruppenhofer et al., 2006) has created a lexical resource for English that documents the range of semantic and syntactic combinatorial possibilities of words and their senses. FrameNet consists of schematic representations of situations (called *frames*), e.g., the frame *win_prize* describes an awarding situation with*frame elements* (FE), i.e., semantic roles, like COMPETITOR, PRIZE, COMPETITION etc.

A pair of a word and a frame forms a *lexical unit* (LU), similar to a particular word sense in a thesaurus. LUs are connected to *lexical entries* (LEs), which capture the valency patterns of frames, providing information about FEs and their phrase types and grammatical functions in relation to the LUs. In total, the FrameNet release 1.5 contains 1019 frames, 9385 lemmas, 11829 lexical units and more than 170,000 annotated sentences.

Comparison to sar-graphs Sar-graphs resemble frames in many aspects, e.g., both define semantic roles for target concepts and provide detailed valency information for linguistic constructions referring to the concept. Table 1 compares some properties of the two resources.

Sar-graphs model relations derived from factual knowledge bases like DBpedia (Lehmann et al., 2015), whereas FrameNet is based on the

FrameNet: A frame …	A **sar-graph** …
… is based on the linguistic theory of frame semantics.	… is defined by a relation in a world-knowledge database.
… groups expressions implicating a situational concept by subsumption.	… groups linguistic structures expressing or implying a relation.
… groups lemmas and their valency patterns.	… groups phrase patterns.
… can have relations to other frames.	… is not explicitly connected to other sar-graphs.

Table 1: Comparison of FrameNet frames to sar-graphs on a conceptual level.

linguistic theory of *frame semantics* (Fillmore, 1976). This theory assumes that human cognitive processing involves an inventory of explicit schemata for classifying, structuring and interpreting experiences. Consequently, FrameNet contains a number of very generic frames (e.g., *forming_relationships*) that have no explicit equivalent in a sar-graph relation. The database-driven sar-graphs also specify fewer semantic roles than frames typically do, covering mainly the most important aspects of a relational concept from a knowledge-base population perspective. For example, the sar-graph for *marriage* lists arguments for the SPOUSES, LOCATION and DATE of the wedding ceremony as well as a DIVORCEDATE, while the related frame *forming_relationships* additionally covers, e.g., an EXPLANATION (divorce reason, etc.) and an ITERATION counter (for the relationships of a person).

Above that, FrameNet specifies relations between frames (*inheritance*, *subframe*, *perspective on*, *using*, *causative of*, *inchoative of*, *see also*) and connects in this way also the lexical units evoking the related frames. For example, frames *commerce_buy* and *commerce_sell* represent perspectives on the frame *commerce_good_transfer*, and link by the same relation the verbs `to sell` and `to buy`. Sar-graphs are currently not linked to one another.

Another difference is the relationship between lexical items and their corresponding frames/sar-graph relations. LUs in FrameNet imply frames by subsumption, e.g., `to befriend` and `to divorce` are subsumed by *forming_relationships*. In comparison, sar-graphs cluster both expressions that directly refer to instances of the target relation (e.g., `to wed` for *marriage*) and those that only entail them (e.g., `to divorce` for *marriage*). This entailment is, in turn, partly represented in FrameNet via frame-to-frame relations like *inheritance*, *cause* and *perspective*.

The data perspective Not only do frames and sar-graphs model different (but related) aspects of the same semantic concepts, they also cover different sets of lexical items, i.e. lemmas with corresponding senses and valency patterns. For example, FrameNet 1.5 neither contains the idiomatic phrase `exchange vows` nor the lemma `remarry` for the *forming_relationships* frame, in contrast to the *marriage* sar-graph; while the sar-graph does not contain all the valency patterns of the LU `widow` which the corresponding frame provides.

A statistical analysis shows that the *marriage* sar-graph and the frames *forming_relationships, personal_relationship, social_connection*, and *relation_between_individuals* share only 7% of their lemmas. The sar-graph adds 62% of the total number of lemmas, FrameNet the remaining 31%. For the *acquisition* relation between companies, values are similar: 6% shared, 79% additional lemmas in the sar-graph, and 15% of the relevant lemmas are only contained in FrameNet.

Linking sar-graphs to FrameNet The similarities between FrameNet and sar-graphs can be used to link the two resources at the level of:

- lexical items (or senses),
- valency patterns and phrase patterns,
- frames and sar-graph relations.

The linking of sar-graphs on the lemma level was already presented in Section 3; in the following we briefly outline some ideas for the (semi-) automatic alignment on the other two levels.

A first linking approach can be to define a similarity metric between sar-graph phrase patterns and FrameNet valency patterns. The metric might include a wide range of semantic and syntactic features of the pattern elements, such as lemma, part of speech, phrase type, grammatical function, and conceptual roles. As both resources work with different label inventories, this would require a manual mapping step on the conceptual level.

	FrameNet	**SarGraph**
lemma	`marry`	`marry`
part of speech	verb	verb, past tense
semantic role	Partner1	Spouse1
role filler	nominal phrase	person mention
gramm. function	external argument	nominal subject
semantic role	Partner2	Spouse2
role filler	nominal phrase	person mention
gramm. function	object	direct object
semantic role	Time	Date
role filler	prep. phrase	date mention
gramm. function	dependent	prep. modifier

Table 2: Example for pattern-level mapping between FrameNet (a valence pattern of LU *marry.v*) and sar-graphs (pattern *marriage*#5088).

However, the effort for this step would be reasonably low because the overall number of labels is relatively small. Table 2 presents an example mapping for patterns covering phrases like "Spouse1 `married` Spouse2 `on` Date".

The described approach can be extended by incorporating annotated sentences from FrameNet which match particular sar-graph patterns, thereby connecting these to the sentences' corresponding valency patterns. The pattern matching can be done automatically, using the same algorithm as when applying patterns to extract novel relation instances from text. Because there are cases where such a match might be misleading (e.g., for long sentences with several mentioned relations), additionally applying a similarity function seems reasonable.

Linking sar-graphs to valency patterns in FrameNet also provides connections on the relation-to-frame level, as every valency pattern is derived from a lexical unit associated with a unique frame. Because of the conceptual differences between FrameNet and sar-graphs, the mapping of frames to relations is not one-to-one but rather a many-to-many linking. For example, the relation *marriage* might equally likely be mapped to one of the more abstract frames *forming_relationships* and *personal_relationship*. The frame *personal_relationships* is related to *personal_relationship* by the inter-frame relation *inchoative of*. The frame *leadership* can be linked to the sar-graph relations *organization leadership* and *organization membership*, since the last one includes also patterns with the lemma `lead` or `leader`, which imply the membership in some group.

Relation	|*Patterns*|	|*V*|	|*E*|
award honor	510	303	876
award nomination	392	369	1,091
country of nationality	560	424	1,265
education	270	233	631
marriage	451	193	584
person alternate name	542	717	1,960
person birth	151	124	319
person death	306	159	425
person parent	387	157	589
person religion	142	196	420
place lived	329	445	1,065
sibling relationship	140	103	260
acquisition	224	268	676
business operation	264	416	876
company end	465	714	1,909
company product rel.	257	421	929
employment tenure	226	131	374
foundation	397	231	708
headquarters	273	220	570
org. alternate name	280	283	720
organization leadership	547	213	717
organization membership	291	262	718
organization relationship	303	317	862
organization type	264	566	1,168
sponsorship	336	523	1,298
Total	**8,307**	**7,988**	**21,010**

Table 3: Dataset statistics

5 Sar-graph dataset

We generated a dataset of sar-graphs for 25 relations from the domains of biographical, awards and business information, with English as the target language. The dataset is available at `http://sargraph.dfki.de`. In this section, we briefly describe some implementation details of the generation process, and present key dataset statistics.

Sar-graph construction We construct sar-graphs using an approach that is language- and relation-independent, and relies solely on the availability of a set of seed relation instances from an existing knowledge base (KB). As described in Section 2, each sar-graph is the result of merging a set of dependency constructions, or patterns. We obtain these dependency constructions by implementing a distantly supervised pattern extraction approach (Mintz et al., 2009; Krause et al., 2012; Gerber and Ngomo, 2014).

We use *Freebase* (Bollacker et al., 2008) as our KB, and select relations of arity $2 \leq n \leq 5$, based on their coverage in Freebase (see Table 3). The selection includes kinship relations (e.g., *mar-*

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<lemon:Lexicon rdf:about="http://dare.dfki.de/lemon/lexicon"
               xmlns:lemon="http://www.monnet-project.eu/lemon#">
<lemon:language>en
<lemon:entry>
  <lemon:LexicalEntry rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024">
    <lemon:canonicalForm>
      <lemon:Form rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#form">
        <lemon:writtenRep xml:lang="en">marry\VBN C_person C_person in\IN C_location
    <lemon:phraseRoot>
      <lemon:Node rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#phraseRoot">
        <root xmlns="http://dare.dfki.de/lemon/ontology#">
          <lemon:Node rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#node1">
            <prep>
              <lemon:Node rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#node4">
                <pobj>
                  ...
                <lemon:leaf>
                  ...
            <lemon:leaf>
              <lemon:Component rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#comp1">
                <lemon:element>
                  <lemon:LexicalEntry rdf:about="http://dare.dfki.de/lemon/lexicon/marry#12024">
                    <lemon:sense>
                      <lemon:LexicalSense rdf:about="http://babelnet.org/synset?word=bn:00090675v"/>
                    <lemon:canonicalForm>
                      <lemon:Form rdf:about="http://dare.dfki.de/lemon/lexicon/marry#form_12024">
                        <lemon:writtenRep xml:lang="en">marry
                    ...
            ...
  <lemon:synBehavior>
    <lemon:Frame rdf:about="http://dare.dfki.de/lemon/lexicon/marriage_12024#frame">
      <person rdf:resource="http://dare.dfki.de/lemon/lexicon/marriage_12024#C_person"
              xmlns="http://dare.dfki.de/lemon/ontology#"/>
      ...
...
```

Figure 3: Excerpt from the sar-graph pattern for the phrase "SPOUSE1 *and* SPOUSE2 *got married in* LOCATION on DATE." In Lemon format; closing tags omitted for brevity.

riage, *parent-child*, *siblings*) and biographical information (*person birth/death*), but also typical inter-business relations and properties of companies (e.g., *acquisition*, *business operation*, *headquarters*). Using Freebase' query API, we retrieved a total of 223K seed instances for the 25 target relations.

The seeds are converted to web search engine queries to generate a text corpus containing mentions of the seeds. We collected a total of 2M relevant documents, which were preprocessed using a standard NLP pipeline for sentence segmentation, tokenization, named entity recognition and linking, lemmatization, part-of-speech tagging and word sense disambiguation. We also applied a dependency parser to annotate sentences with Stanford dependency relations. After preprocessing, we discarded duplicate sentences, and sentences that did not contain mentions of the seed relation instances.

From the remaining 1M unique sentences, we extracted 600K distinct dependency patterns by finding the minimum spanning tree covering the arguments of a given seed instance. To reduce the number of low-quality patterns, a side effect of the distantly supervised learning scheme, we implemented the filtering strategies proposed by Moro et al. (2013). These strategies compute confidence metrics based on pattern distribution statistics and on the semantic coherence of a pattern's content words. Patterns with low confidence scores are discarded. To create a sar-graph instance, we then merge the patterns based on their shared relation argument vertices (see Figure 1). Sar-graph instances, patterns, and vertices are assigned unique ids to support efficient lookup.

Dataset statistics and format Table 3 summarizes key statistics of the dataset. The curated sar-graphs range in size from 140–560 unique patterns. The largest sar-graph, for the *person alternate name* relation, contains 1960 edges and 717 vertices. The smallest sar-graph was constructed for the *sibling* relation, it contains 260 edges and 103 vertices, derived from 140 dependency patterns. Overall, the dataset contains approximately 8,300 unique patterns. While this experimental dataset is not as large as other linguistic LOD resources, we emphasize that the construction of additional sar-graph instances, e.g., for other relations or a different language, is a fully automatic process given a set of seed relation instances.

We provide the dataset in a custom, XML-based format, and in the semantic web dialect *Lemon*.[1] Lemon was originally designed for modeling dic-

[1] http://www.lemon-model.net/

tionaries and lexicons. It builds on RDF and provides facilities for expressing lexicon-relevant aspects of a resource, e.g., lexical items with different forms and senses. Albeit Lemon is not a perfect fit for representing sar-graphs and their individual pattern elements, it still constitutes a good first step for establishing sar-graphs as part of the linguistic linked open data cloud.

Figure 3 shows an example pattern in Lemon format. Patterns are realized via Lemon *lexicon entries*, where each such entry has an attached phrase root whose child nodes contain information about the syntactic and lexical elements of the pattern.

Java-based API We provide a Java-based API which simplifies loading, processing, and storing sar-graphs. One exemplary API feature are materialized views, which present the sar-graph data in the respective most informative way to an application, as with different tasks and goals, varying aspects of a sar-graph may become relevant.

6 Related Work

In comparison to well-known knowledge bases such as YAGO (Suchanek et al., 2008), DBpedia (Lehmann et al., 2015), Freebase (Bollacker et al., 2008), or the recent Google Knowledge Vault (Dong et al., 2014), sar-graphs are not a database of facts or events, but rather a repository of linguistic expressions of these. The acquisition of sar-graph elements is related to pattern discovery approaches developed in traditional schema-based IE systems, e.g., NELL (Mitchell et al., 2015) or PROSPERA (Nakashole et al., 2011), meaning that sar-graphs can be directly applied to free texts for enlarging a structured repository of knowledge.

Many linguistic resources, such as WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998), and VerbNet (Schuler, 2005) already existed before the recent development of large knowledge bases. These resources model the semantics of languages at the word or syntactic level, without an explicit link to real world facts. Most of them were manually created and are relatively small. WordNet captures lexical semantic relations between individual words, such as synonymy, homonymy, and antonymy. FrameNet focuses on fine-grained semantic relations of predicates and their arguments. VerbNet is a lexicon that maps verbs to predefined classes which define the syntactic and semantic preferences of the verb. In contrast to these resources, sar-graphs are data-driven, constructed automatically, and incorporate statistical information about relations and their arguments. Therefore, sar-graphs complement these manually constructed linguistic resources.

There is also increasing research in creating large-scale linguistic resources, e.g., BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer and Havasi, 2013) and UBY (Gurevych et al., 2012) automatically. Many of these are built on top of existing resources like WordNet, Wiktionary and Wikipedia, e.g., BabelNet merges Wikipedia concepts including entities with word senses from WordNet. ConceptNet is a semantic network encoding common-sense knowledge and merging information from various sources such as WordNet, Wiktionary, Wikipedia and ReVerb. In comparison to sar-graphs, it contains no explicit linguistic knowledge like syntactic or word-sense information assigned to the content elements, and the semantic relations among concepts are not fixed to an ontology or schema. UBY combines and aligns several lexico-semantic resources, and provides a standardized representation via the Lexical Markup Framework.

7 Conclusion

We presented sar-graphs, a linguistic resource linking semantic relations from knowledge graphs to their associated natural language expressions. Sar-graphs can be automatically constructed for any target language and relation in a distantly supervised fashion, i.e. given only a set of seed relation instances from an existing knowledge graph, and a text corpus. We publish an initial dataset which contains sar-graphs for 25 Freebase relations, spanning the domains of biographical, award, and business information. The released sar-graphs are linked at the lexical level to BabelNet, WordNet and UBY, and are made available in Lemon-RDF and a custom XML-based format at `http://sargraph.dfki.de`.

For future releases of the sar-graph dataset, we intend to publish the non-curated part of the pattern data, and to provide more detailed information about the source of linguistic expressions (i.e., to expand the public data with source sentences and seed facts). Furthermore, we will continue our work on linking sar-graphs to FrameNet, in particular we will focus on semi-automatic phrase-level linking, for which we have outlined some early ideas in this paper. We also plan to expand the dataset to more relations and additional languages.

Acknowledgments

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the projects ALL SIDES (01IW14002) and BBDC (01IS14013E), by the German Federal Ministry of Economics and Energy (BMWi) through the projects SDW (01MD15010A) and SD4M (01MD15007B), and by Google through a Focused Research Award granted in July 2013. We would also like to thank Min Fang and Hong Li for their help with the implementation.

References

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proc. of ACL-COLING*, pages 86–90.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proc. of SIGMOD*, pages 1247–1250.

G. de Melo and G. Weikum. 2009. Towards a Universal Wordnet by Learning from Combined Evidence. In *Proc. of CIKM*, pages 513–522.

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proc. of SIGKDD*, pages 601–610.

Ch. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

C. J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Daniel Gerber and Axel-Cyrille Ngonga Ngomo. 2014. From RDF to natural language and back. In *Towards the Multilingual Semantic Web*. Springer Berlin Heidelberg.

I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, Ch. M. Meyer, and Ch. Wirth. 2012. Uby: A Large-scale Unified Lexical-semantic Resource Based on LMF. In *Proc. of EACL*, pages 580–590.

S. Krause, H. Li, H. Uszkoreit, and F. Xu. 2012. Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web. In *Proc. of ISWC*, pages 263–278.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and Ch. Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of ACL/IJCNLP*, pages 1003–1011.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proc. of AAAI*.

A. Moro, H. Li, S. Krause, F. Xu, R. Navigli, and H. Uszkoreit. 2013. Semantic Rule Filtering for Web-Scale Relation Extraction. In *Proc. of ISWC*, pages 347–362.

A. Moro, A. Raganato, and R. Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *TACL*, 2:231–244.

N. Nakashole, M. Theobald, and G. Weikum. 2011. Scalable Knowledge Harvesting with High Precision and High Recall. In *Proc. of WSDM*, pages 227–236.

R. Navigli and S. P. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

K. K. Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

R. Speer and C. Havasi. 2013. ConceptNet 5: A Large Semantic Network for Relational Knowledge. In *The People's Web Meets NLP*, pages 161–176. Springer Berlin Heidelberg.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

H. Uszkoreit and F. Xu. 2013. From Strings to Things – Sar-Graphs: A New Type of Resource for Connecting Knowledge and Language. In *Proc. of WS on NLP and DBpedia*.