A Distributional Semantic Search Infrastructure for Linked Dataspaces

André Freitas, Seán O'Riain, and Edward Curry

Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway

Abstract. This paper describes and demonstrates a distributional semantic search service infrastructure for Linked Dataspaces. The center of the approach relies on the use of a distributional semantics infrastructure to provide semantic search and query services over data for users and applications, improving data accessibility over the Dataspace. By accessing the services through a REST API, users can semantically index and search over data using the distributional semantic knowledge embedded in the reference corpus. The use of distributional semantic models, which rely on the automatic extraction from large corpora, supports a comprehensive and approximative semantic matching mechanism with a low associated adaptation cost for the inclusion of new data sources.

Keywords: Distributional Semantics, Semantic Matching, Semantic Search, Explicit Semantic Analysis, Dataspaces, Linked Data.

1 Motivation

Within the realm of the Web and of Big Data, dataspaces where data is more complex, sparse and heterogeneous are becoming more common. Consuming this data demands applications and search/query mechanisms with the semantic flexibility necessary to cope with the semantic/vocabulary gap between users, different applications and data sources within the dataspace. Traditionally, consuming structured data demands that users, applications and databases share the same vocabulary before data consumption, where the semantic matching process is done manually. As dataspaces grow in complexity, the ability to semantically search over data using one's own vocabulary becomes a fundamental functionality for dataspaces.

Distributional semantics [1, 2] applied to semantic search can become a fundamental element in enabling semantic search and semantic matching in dataspaces, by providing a comprehensive and low-cost semantic model based on unstructured data.

This paper demonstrates a distributional semantics service infrastructure for Linked Dataspaces. The demonstration is exemplified over a question answering and a vocabulary search scenarios using DBpedia as the data source. In addition to these scenarios, distributional semantic search over dataspaces can be used for tasks such as ontology alignment, service matching, content recommendation, entity disambiguation, among others.

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7955, pp. 214–218, 2013.

2 Distributional Semantic Infrastructure

Distributional semantics is defined upon the assumption that the context surrounding a given word in a text provides important information about its meaning [2]. It focuses on the construction of a simplified semantic model for a word based on the statistical distribution of co-occurring words in large text collections. These semantic models are naturally represented by Vector Space Models (VSMs) [2], where the meaning of a word can be defined by a weighted vector over a term or concept space, which represents the association pattern of co-occurring words in a reference corpora. The existence of large amounts of unstructured text on the Web brings the potential to create comprehensive distributional semantic models (DSMs). DSMs can be automatically built from large corpora, not requiring manual intervention on the creation of the semantic model. These characteristics facilitates the creation of DSMs for multiple languages and can provide a more scalable solution to the problem of capturing large-scale commonsense semantic information, necessary for providing a flexible semantic model that could bridge the semantic gap between users, applications and data.

The suitability of the application of distributional semantics to semantic search is empirically supported [1, 2]. Distributional indexes and vector space abstractions (T-Space) specific for Entity-Attribute-Value (EAV) and graph databases were also formulated [2]. These recent developments bring the potential for distributional semantic models and distributional data indexes to become recurrent infrastructure elements for dataspaces. This paper demonstrates different services provided by a distributional semantic infrastructure inside the *Treo* framework, where a set of semantic operations over dataspaces are used to provide a natural language/semantic search interface over DBpedia.

3 High-Level Architecture Components

Figure 1 shows the high-level architecture components of the *Treo* distributional semantic infrastructure. The functionalities of the components are described below.

- **Distributional-Compositional Semantic Index:** Inverted index which implements the T-Space structured vector space model (VSM) [2]. The T-Space VSM allows a principled distributional-compositional semantic representation of labelled data graphs, supporting search operations over the data graphs with semantic approximation.
- **Explicit Semantic Analysis (ESA):** Provides the distributional concept vectors based on the TF/IDF indexing of the reference corpora (in this example English Wikipedia). Each word present in the graph is represented in the index by a weighted concept vector based on the ESA semantic interpretation approach.
- **Distributional Search:** Distributional search operations allow a semantic approximation/matching between the query terms and the graph triple labels

(*subject*, *predicate* and *object*). Different search patterns can be employed over the data graph. In the natural language query mechanism the query is translated into a set of graph search operations on the distributional index.
- **Instance Search:** Keyword-based instance search which ranks instances according to instance cardinality, string similarity and uses disambiguation/sameAs links when possible (non-distributional search).
- **Distributional Indexer:** Indexes a graph with a labelled graph/EAV representation over the distributional-compositional index (T-Space).
- **Semantic Relatedness:** Computes the ESA semantic relatedness measure between two words. Distributional semantic relatedness measures can be used in tasks such as entity disambiguation, ontology/schema alignment, among others.

The core service components can be accessed by third-party applications through a RESTFul web API.

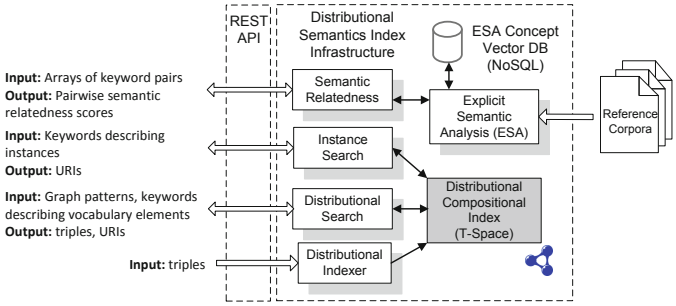

Fig. 1. High-level components of the *Treo* distributional semantics infrastructure

4 Demonstration

In this demonstration the core services available on the distributional semantic infrastructure are used by the components of the natural language interface (NLI) of the *Treo* query engine. Two query types are supported: (i) free natural language queries over graph data (e.g. *'How tall is Claudia Schiffer?'*) or (ii) vocabulary queries over graph elements (e.g. *give me all predicates related to the concept 'geology'*). The second query type is fundamental for tasks such as exploratory dataset search, ontology alignment and matching as well as in the search and reuse existing schema elements. Both queries answer sets are displayed in Figure 2.

The *distributional indexer* service was used to index DBpedia 3.7. Taking the natural language query *'How tall is Claudia Schiffer?'*, the *Treo* NLI system parses the natural language query, transforming it into a triple-like representation. To determine the core entity of the query, the entity resolution module uses

the *instance search* and the *distributional search* services to search for instances and classes that match in the dataset. The named entity *'Claudia Schiffer'* is defined as the pivot query entity and mapped to the URI *dbpedia:Claudia_Schiffer*. The query triple representation *dbpedia:Claudia_Schiffer - tall* is then determined and this query pattern is sent to the *distributional search* service for the predicate semantic matching. Using the distributional representation of the indexed DBpedia graph, the query term *'tall'* is matched against *dbpedia-owl:height* predicate for Claudia Schiffer and the triple pattern *dbpedia:Claudia_Schiffer - dbpedia-owl:height* is sent as a graph pattern to the *distributional search* service, which returns the result in Figure 2. In the second example, the user wants the predicates which are related to the *geology* domain inside DBpedia. The keyword *'geology'* is sent to the *distributional search* service which returns a list of semantically related predicates URIs (Figure 2).

The examples show how the distributional infrastructure uses the semantic knowledge extracted from the reference corpora to semantically search and rank datasets' elements. No additional ontological information is used for the semantic matching. Distributional search queries can be translated as the computation and ranking of semantic relatedness scores between query terms and dataset elements. The reader can find additional query examples in the website[1] and additional information on distributional semantic search, including the evaluation under the Question Answering over Linked Data test collection[2] in [1,2].

5 Related Applications

PowerMap [3] is a hybrid matching algorithm which includes terminological and structural schema matching techniques with the assistance of large scale ontological and lexical resources. PowerMap uses WordNet based similarity approaches as a semantic approximation strategy. PowerAqua, a question answering system for Linked Data, uses PowerMap to match query terms to vocabulary terms. The Semantic Web Search Engine (SWSE) [5] is a search and query service that implements an architecture with components for crawling, integrating, indexing, querying, and navigating over multiple data sources. The systems main components include query processing, ranking, an index manager, and an internal data store. Sindice [4] is a search and query service for the Linked Data Web that ranks entities according to the incidence of keywords associated with them. It uses a node-labelled tree model to represent the relationship between datasets, entities, attributes, and values. Similarly to SWSE, Sindice provides a comprehensive entity-centric search and query infrastructure. Compared to PowerMap, the distributional infrastructure for Treo concentrates its semantic matching strategy on DSMs, implementing the DSM as a distributional-compositional (T-Space) semantic index. Compared to SWSE and Sindice, the *Treo* semantic index focuses on the problem of the semantic matching strategies, while SWSE and Sindice does not focus on the vocabulary problem.

[1] `http://treo.deri.ie/eswcdemo`
[2] `http://www.sc.cit-ec.uni-bielefeld.de/qald-1`

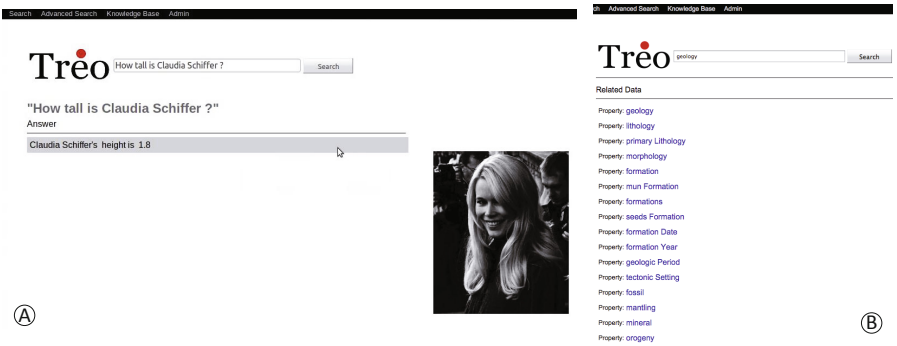

Fig. 2. Answer sets for the query examples: (A) Natural language query over the DBpedia data graph and (B) Keyword-based semantic search over the dataset terminological level

6 Conclusions and Future Work

This work demonstrates the Treo distributional semantic infrastructure for Linked Dataspaces using natural language queries and terminological search operations over DBpedia. Future work will concentrate on improvements over the distributional semantic model.

References

1. Freitas, A., Curry, E., O'Riain, S.: A Distributional Approach for Terminological Semantic Search on the Linked Data Web. In: Proc. of the 27th ACM Symposium on Applied Computing (SAC), Semantic Web and Applications, SWA (2012)
2. Freitas, A., Curry, E., Oliveira, J.G., O'Riain, S.: A Distributional Structured Semantic Space for Querying RDF Graph Data. International Journal of Semantic Computing, IJSC (2012)
3. Lopez, V., Sabou, M., Motta, E.: PowerMap: Mapping the Real Semantic Web on the Fly. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 414–427. Springer, Heidelberg (2006)
4. Delbru, R., Campinas, S., Tummarello, G.: Searching Web Data: An Entity Retrieval and High-Performance Indexing Model. J. Web Semantics (2011)
5. Hogan, A., et al.: Searching and Browsing Linked Data with SWSE: The Semantic Web Search Engine. J. Web Semantics (2011)