

# Towards a Distributional Semantic Web Stack

André Freitas<sup>1</sup>, Edward Curry<sup>1</sup>, Siegfried Handschuh<sup>1,2</sup>

<sup>1</sup>Insight Centre for Data Analytics, National University of Ireland, Galway

<sup>2</sup>School of Computer Science and Mathematics, University of Passau

**Abstract.** The capacity of distributional semantic models (DSMs) to discover similarities over large scale heterogeneous and poorly structured data brings them as a promising universal and low-effort framework to support semantic approximation and knowledge discovery. This position paper explores the role of distributional semantics in the Semantic Web vision, based on state-of-the-art distributional-relational models, categorizing and generalizing existing approaches into a Distributional Semantic Web stack.

## 1 Introduction

*Distributional semantics* is based on the idea that semantic information can be extracted from lexical co-occurrence from large-scale data corpora. The simplicity of its vector space representation, its ability to automatically derive meaning from large-scale unstructured and heterogeneous data and its built-in semantic approximation capabilities are bringing distributional semantic models as a promising approach to bring additional flexibility into existing knowledge representation frameworks.

Distributional semantic approaches are being used to complement the semantics of structured knowledge bases, generating hybrid *distributional-relational models*. These hybrid models are built to support *semantic approximation*, and can be applied to selective reasoning mechanisms, reasoning over incomplete KBs, semantic search, schema-agnostic queries over structured knowledge bases and knowledge discovery.

## 2 Distributional Semantic Models

*Distributional semantic models* (DSMs) are semantic models which are based on the statistical analysis of co-occurrences of words in large corpora. Distributional semantics allows the construction of a *quantitative model of meaning*, where the degree of the semantic association between different words can be quantified in relation to a *reference corpus*. With the availability of large Web corpora, comprehensive distributional models can effectively be built.

DSMs are represented as a *vector space model*, where each dimension represents a *context*  $\mathcal{C}$  for the linguistic or data context in which the *target term*  $\mathcal{T}$  occurs. A *context* can be defined using documents, co-occurrence window sizes

(number of neighboring words or data elements) or syntactic features. The *distributional interpretation* of a target term is defined by a weighted vector of the contexts in which the term occurs, defining a geometric interpretation under a distributional vector space. The weights associated with the vectors are defined using an *associated weighting scheme*  $\mathcal{W}$ , which can re-calibrates the relevance of more generic or discriminative contexts. A *semantic relatedness measure*  $\mathcal{S}$  between two words in the dataset can be calculated by using different *similarity/distance* measures such as the *cosine similarity* or *Euclidean distance*. As the dimensionality of the distributional space can grow large, dimensionality reduction approaches  $d$  can be applied.

Different DSMs are built by varying the parameters of the tuple  $(\mathcal{T}, \mathcal{C}, \mathcal{W}, d, \mathcal{S})$ . Examples of distributional models are *Latent Semantic Analysis*, *Random Indexing*, *Dependency Vectors*, *Explicit Semantic Analysis*, among others. Distributional semantic models can be specialized to different application areas using different corpora.

### 3 Distributional-Relational Models (DRMs)

*Distributional-Relational Models* (DRMs) are models in which the semantics of a *structured knowledge base* (KB) is complemented by a *distributional semantic model*.

A *Distributional-Relational Model* (DRM) is a tuple  $(\mathcal{DSM}, \mathcal{KB}, \mathcal{RC}, \mathcal{F}, \mathcal{OP})$ , where:  $\mathcal{DSM}$  is the *associated distributional semantic model*;  $\mathcal{KB}$  is the *structured dataset*, with elements  $E$  and tuples  $\Omega$ ;  $\mathcal{RC}$  is the *reference corpora* which can be unstructured, structured or both. The reference corpora can be internal (based on the co-occurrence of elements within the  $\mathcal{KB}$ ) or external (a separate reference corpora);  $\mathcal{F}$  is a *map* which translates the elements  $e_i \in E$  into vectors  $\vec{e}_i$  in the the distributional vector space  $VS^{\mathcal{DSM}}$  using the natural language label and the entity type of  $e_i$ ;  $\mathcal{OP}$  is a set of *operations* over  $\vec{e}_i$  in  $VS^{\mathcal{DSM}}$  and over  $E$  and  $\Omega$  in the  $\mathcal{KB}$ . The set of operations may include *search*, *query* and *graph navigation* operations using the distance measure  $\mathcal{S}$ .

The DRM supports a double perspective of semantics, keeping the fine-grained precise semantics of the structured  $KB$  but also complementing it with the distributional model. Two main categories of DRMs and associated applications can be distinguished:

**Semantic Matching & Commonsense Reasoning:** In this category the  $\mathcal{RC}$  is unstructured and it is distinct from the  $\mathcal{KB}$ . The large-scale *unstructured*  $\mathcal{RC}$  is used as a *commonsense knowledge base*. Freitas & Curry [1] define a DRM ( $\tau - Space$ ) for supporting schema-agnostic queries over the structured  $\mathcal{KB}$ : terms used in the query are projected into the distributional vector space and are semantically matched with terms in the  $\mathcal{KB}$  via distributional semantics using commonsense information embedded on large scale unstructured corpora  $\mathcal{RC}$ . In a different application scenario, Freitas et al. [3] uses the  $\tau - Space$  to support selective reasoning over commonsense  $\mathcal{KB}$ s. Distributional semantics is used to select the facts which are semantically relevant under a specific reasoning

context, allowing the scoping of the reasoning context and also coping with incomplete knowledge of commonsense  $KBs$ . Pereira da Silva & Freitas [2] used the  $\tau - Space$  to support approximate reasoning on logic programs.

**Knowledge Discovery:** In this category, the structured  $\mathcal{KB}$  is used as a distributional reference corpora (where  $\mathcal{RC} = \mathcal{KB}$ ). Implicit and explicit semantic associations are used to derive new meaning and discover new knowledge. The use of structured data as a distributional corpus is a pattern used for knowledge discovery applications, where knowledge emerging from *similarity patterns in the data* can be used to retrieve similar entities and expose implicit associations. In this context, the ability to represent the  $\mathcal{KB}$  entities' attributes in a vector space and the use of vector similarity measures as way to retrieve and compare similar entities can define universal mechanisms for knowledge discovery and semantic approximation. Novacek et al. [5] describe an approach for using web data as a bottom-up phenomena, capturing meaning that is not associated with explicit semantic descriptions, applying it to entity consolidation in the life sciences domain. Speer et al. [8] proposed AnalogySpace, a DRM over a commonsense  $\mathcal{KB}$  using Latent Semantic Indexing targeting the creation of the analogical closure of a semantic network using dimensional reduction. AnalogySpace was used to reduce the sparseness of the  $\mathcal{KB}$ , generalizing its knowledge, allowing users to explore implicit associations. Cohen et al. [6] introduced PSI, a predication-based semantic indexing which uses a structured biomedical  $\mathcal{KB}$  also as a distributional reference corpus. PSI was used for similarity-based retrieval and detection of implicit associations.

## 4 The Distributional Semantic Web Stack

DRMs provide universal mechanisms which have fundamental features for semantic systems: (i) built-in semantic approximation for terminological and instance data; (ii) ability to use large-scale unstructured data as commonsense knowledge, (iii) ability to detect emerging implicit associations in the  $\mathcal{KB}$ , (iv) simplicity of use supported by the vector space model abstraction, (v) robustness with regard to poorly structured, heterogeneous and incomplete data. These features provide a framework for a robust and easy-to-deploy semantic approximation component grounded on large-scale data. Considering the relevance of these features in the deployment of semantic systems in general, this paper synthesizes its vision by proposing a *Distributional Semantic Web stack* abstraction (Figure 1), complementing the Semantic Web stack. At the bottom of the stack, unstructured and structured data can be used as reference corpora together with the target  $\mathcal{KB}$  (RDF(S)). Different elements of the distributional model are included as optional and composable elements of the architecture. The *approximate search and query operations layer* access the *DSM layer*, supporting users with semantically flexible search and query operations. A *graph navigation layer* defines graph navigation algorithms (e.g. such as spreading activation, bi-directional search) using the semantic approximation and the distributional information from the layers below.

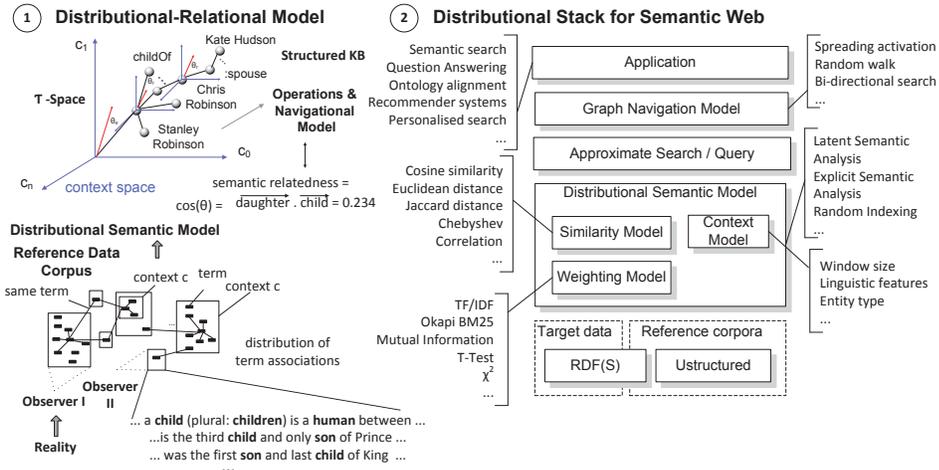


Fig. 1: (A) Depiction of an example DRM ( $\tau$ -Space) (B) Distributional Semantic Web stack.

**Acknowledgment:** This work has been joint funded by the Irish Research Council.

## References

- Freitas, A., Curry, E., Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional-Compositional Semantics Approach. *In Proc. of the 19th Intl. Conf. on Intelligent User Interfaces (IUI)*. (2014).
- Pereira da Silva, J.C., Freitas A., Towards An Approximative Ontology-Agnostic Approach for Logic Programs, *In Proc. of the 8th Intl. Symposium on Foundations of Information and Knowledge Systems*. (2014).
- Freitas, A., Pereira Da Silva, J.C., Curry, E., Buitelaar, P., A Distributional Semantics Approach for Selective Reasoning on Commonsense Graph Knowledge Bases. *In Proc. of the 19th Int. Conf. on Applications of Natural Language to Information Systems (NLDB)*. (2014).
- Speer, R., Havasi, C., Lieberman, H., AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. *In Proc. of the 23rd Intl. Conf. on Artificial Intelligence*, 548-553. (2008).
- Novacek, V., Handschuh, S., Decker, S.. Getting the Meaning Right: A Complementary Distributional Layer for the Web Semantics. *In Proc. of the Intl. Semantic Web Conference*, 504-519. (2011).
- Cohen, T., Schvaneveldt, R.W., Rindfleisch, T.C.. Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space. *T. AMIA Annu Symp Proc.*, 114-118. (2009).
- Turney, P.D., Pantel P., From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1), 141-188. (2010).
- Speer, R., Havasi, C., Lieberman, H., AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. *In Proc. of the 23rd Intl. Conf. on Artificial Intelligence*, 548-553. (2008).