Distributional-Relational Models: Scalable Semantics for Databases

André Freitas[1,2], **Siegfried Handschuh**[1], **Edward Curry**[2]
[1]Faculty of Computer Science and Mathematics
University of Passau
[2]Insight Centre for Data Analytics
National University of Ireland, Galway

Abstract

The crisp/brittle semantic model behind databases limits the scale in which data consumers can query, explore, integrate and process structured data. Approaches aiming to provide more comprehensive semantic models for databases, which are purely logic-based (e.g. as in Semantic Web databases) have major scalability limitations in the acquisition of structured semantic and commonsense data. This work describes a complementary semantic model for databases which has semantic approximation at its center. This model uses distributional semantic models (DSMs) to extend structured data semantics. DSMs support the automatic construction of semantic and commonsense models from large-scale unstructured text and provides a simple model to analyze similarities in the structured data. The combination of distributional and structured data semantics provides a simple and promising solution to address the challenges associated with the interaction and processing of structured data.

Introduction

Data consumers querying, exploring, integrating or analyzing data today need to go through the process of mapping their own conceptualization to the identifiers of database elements. The requirement of a perfect symbolic and syntactic matching in the database interaction process forces the user to perform (during query construction time) a time consuming *information need-database symbol* alignment process. With the growth of the symbolic space associated with contemporary databases, the process of manual alignment to the database symbolic space becomes infeasible and restrictive.

Automatic semantic approximation between the data consumer information needs and database elements is a central operation for data querying, exploration, integration and data analysis. However, effective semantic approximation is heavily dependent on the construction of comprehensive semantic/commonsense knowledge bases. While different semantic approaches based on logical frameworks have been proposed, such as Semantic Web databases, these approaches are limited in addressing the trade-off between providing an expressive semantic representation and the ability to acquire comprehensive knowledge bases under that representation model. Logical frameworks are highly sensitive to problems from the consistency and from the performance perspectives, which emerge in large-scale knowledge bases.

This work proposes the use of *distributional semantic models* (DSMs) to address these limitations, where the *simplification of the semantic representation* in DSMs facilitates the construction of *large-scale and comprehensive semantic/commonsense knowledge bases*, which *can be used to support effective semantic approximations* for databases. Distributional semantics provides a complementary perspective to the formal perspective of database semantics, which supports *semantic approximation as a first-class database operation*.

A distributional semantics approach implies extending the formal database semantics with a distributional semantic layer. In the hybrid model, the crisp semantics of query terms and database elements are extended and grounded over a distributional semantic model (Figure 1). The distributional layer can be used to abstract the database user from the specific conceptualization of the data.

Distributional Semantics

Distributional semantics is built upon the assumption that the context surrounding a given word in a text provides important information about its meaning (Harris 1954), (Turney and Pantel 2010). Distributional semantics focuses on the construction of a semantic representation of a word based on the statistical distribution of word co-occurrence in unstructured data. The availability of high volume and comprehensive Web corpora brought distributional semantic models as a promising approach to build and represent meaning at scale.

One of the major strengths of distributional models is from the *acquisitional* point of view, where a semantic model can be *automatically built from large unstructured text*. In Distributional Semantic Models (DSMs) the meaning of a word is represented by a weighted vector, which can be automatically built from contextual co-occurrence information in unstructured data. The *distributional hypothesis* (Harris 1954) assumes that the *local context* in which a term occurs can serve as discriminative semantic features which represent the meaning of the term. While this simplification makes distributional semantics a coarse-grained semantic

model, not suitable for all tasks, the scale in which semantic knowledge and associations can be captured makes them effective models for *calculating semantic approximations*, a fact which is supported by empirical evidence (Gabrilovich and Markovitch 2007).

DSMs are represented as a *vector space model*, where each dimension represents a *context pattern* $\mathcal{C}$ for the linguistic or data context in which the *target term* $\mathcal{T}$ occurs. A *context* can be defined using documents, data tuples, co-occurrence window sizes (number of neighboring words) 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 calibrates the relevance of more generic or discriminative contexts. The *semantic relatedness measure* s between two words is calculated by using different *similarity/distance* measures such as the *cosine similarity*, *Euclidean distance*, *mutual information*, among others. As the dimensionality of the distributional space grows, dimensionality reduction approaches d can be applied.

Distributional-Relational Models (DRMs)

The semantics of a database element e (e.g. constants, predicates) is represented by the set of natural language descriptors associated with it. This typically does not include concept associations outside the scope of the specific task that the database was designed to address, limiting its use for semantic approximation to concepts outside the designed database representation. Semantic approximation operations are a fundamental operation to support *schema-agnosticism* (Freitas, Silva, and Curry 2014), i.e. the ability to interact with a database without a precise understanding of the conceptual model behind it.

In this work, the formal semantics of a database symbol is extended with a distributional semantics description, which captures the large-scale symbolic associations within a large reference corpora. The distributional semantics representation captures large-scale semantic, commonsense and domain specific knowledge, using it in the semantic approximation process between a third-party information need and the database (Figure 3). The hybrid distributional-structured model is called *Distributional-Relational Model* (DRM). A DRM embeds the structure defined by relational models in a distributional vector space, where every entity and relationship have an associated vector representation. The distributional associational information embedded in the distributional vector space is used to semantically complement the knowledge expressed in the structured data model. The distributional information is then used to support semantic approximations, while preserving the semantics of the structured data.

A *Distributional-Relational Model (DRM)* is a tuple $(\mathcal{DSM}, \mathcal{DB}, \mathcal{RC}, \mathcal{F}, \mathcal{H})$, where: $\mathcal{DSM}$ is the *associated distributional semantic model*; $\mathcal{DB}$ is the *database* with elements E; $\mathcal{RC}$ 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 DB)

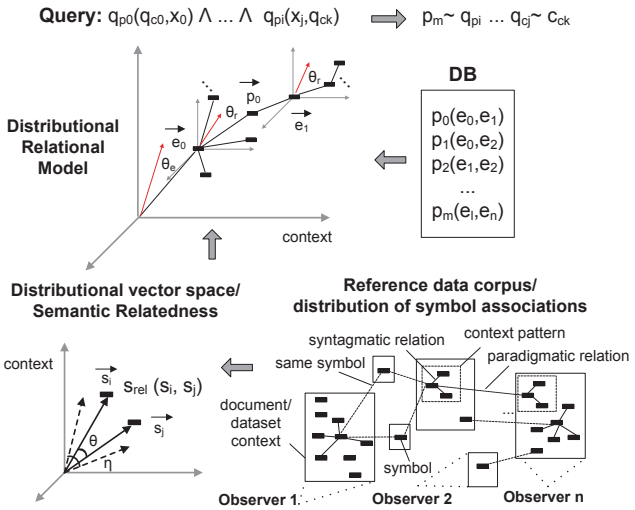

Figure 1: Depiction of distributional relations, contexts and different representation views for distributional semantics.

or external (a separate reference corpora); $\mathcal{F}$ is a *map* which translates the elements $e_i \in E$ into vectors $\overrightarrow{\mathbf{e_i}}$ in the the distributional vector space $VS^{\mathcal{DSM}}$ using the natural language descriptor of e_i; and $\mathcal{H}$ is the set of thresholds above which two terms are semantically equivalent.

Definition (Distributional-Relational Model (DRM)): A distributional relational model is a tuple $(\mathcal{DSM}, \mathcal{DB}, \mathcal{RC}, \mathcal{F}, \mathcal{H})$ such that:

- $\mathcal{DSM}$ is the *associated distributional semantic model*.
- $\mathcal{DB}$ is an *structured dataset* with DB elements E and tuples T.
- $\mathcal{RC}$ 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 DB) or external (a separate reference corpora).
- $\mathcal{F}$ is a *map* which translates the elements $e_i \in E$ into vectors $\overrightarrow{\mathbf{e_i}}$ in the the distributional vector space $VS^{\mathcal{DSM}}$ using the string of e_i and the data model category of e_i.
- $\mathcal{H}$ is the set of *semantic thresholds* for the distributional semantic relatedness s in which two terms are considered semantically equivalent if they are equal and above the threshold.

In this work we assume a simplified data model with a signature $\Sigma = (P, C)$ formed by a pair of finite set of symbols used to represent binary and unary *predicates* $p \in P$ between *constants* $c \in C$. The semantics of the DB is defined by the vectors in the distributional space used to represent the elements. The set of all distributional contexts $Context = \{\chi_1, \cdots, \chi_t\}$ are extracted from a reference corpus and each context $\chi_i \in Context$ is mapped to an identifier which represents the co-occurrence pattern in the corpus. Each identifier χ_i defines a set which tracks the context where a term t mapping to the database element e occurred. This set is used to construct the basis $Context_{base} = \{\overrightarrow{\chi_1}, \cdots, \overrightarrow{\chi_t}\}$ of vectors that spans the *distributional vector space* VS^{dist} (Figure 1). Once the DSM is built, the elements of the signature Σ of the DB are translated into vectors in VS^{dist}. The vector representation of E,

VS^{dist} is defined by:

$$VS^{dist} = \{\overrightarrow{\mathbf{e}} : \overrightarrow{\mathbf{e}} = \sum_{i=1}^{t} w_i^e \overrightarrow{\chi}_i, \text{for each } e \in E\} \quad (1)$$

where w_i^e are defined by a co-occurrence weighting scheme.

The last step refers to the translation of DB tuples into VS^{dist} elements. As each relation and entity symbol has a vector representation, we can define the vector representation of an atom r in the concept vector space by the following definition.

Definition (Distributional Representation of a Tuple): Let $\overrightarrow{\mathbf{p}}$, $\overrightarrow{\mathbf{c_1}}$ and $\overrightarrow{\mathbf{c_2}}$ be the vector representations, respectively, of the binary predicate p and its associated constants c_1 and c_2. A tuple vector representation (denoted by $\overrightarrow{\mathbf{r}}$) is defined by: $(\overrightarrow{\mathbf{p}} - \overrightarrow{\mathbf{c_1}})$ if $p(c_1)$; $(\overrightarrow{\mathbf{p}} - \overrightarrow{\mathbf{c_1}}, \overrightarrow{\mathbf{c_2}} - \overrightarrow{\mathbf{p}})$ if $p(c_1, c_2)$.

Semantic Approximation & Context

The computation of the *distributional semantic relatedness* is a semantic approximation process in which the distributional semantic knowledge and relatedness measure serves as surrogates for the rules, axioms and inference in a deductive logic approach. The assumption is that the knowledge that would be expressed as rules and axioms in a logical commonsense KB is partially embedded in an unstructured way in the reference corpora, and that an *initial query-database alignment provides the contextual and scoping mechanism in which the distributional knowledge can be applied as a semantic/commonsense approximation mechanism.*

The *distributional alignment* (*d-alignment*) $t_q \sim^{\mathcal{DSM}} t_{DB}$ between a query term t_q and a database term t_{DB} is defined when $s_{rel}(t_q, t_{DB}) \geq \eta$. A d-alignment is not equivalent to $t_q \equiv t_{DB}$ in absolute terms, i.e. for all possible inference contexts. However, we argue that given an initial query-database alignment context (a *semantic pivot*), the d-alignment can be locally equivalent to $t_q \equiv t_{DB}$.

Contextual Distributional Equivalence Hypothesis: Let t_q be a query term and t_{DB} be a database term. Let $\kappa(t_q)$ and $\kappa(t_{DB})$ be the contextual information associated with the query and database terms respectively (i.e. previous alignments). If t_q is d-aligned with t_{DB} under the context $(\kappa(t_q), \kappa(t_{DB}))$ then $t_q \equiv t_{DB}$, i.e. can be assumed to be semantically equivalent.

For a semantic approximation process, the context sets should be selected in a way which minimizes the probability of a semantic mismatching using a *heuristic function* which prioritizes the easiest alignment to address, i.e., the query term with *lower semantic entropy* (i.e. less subject to ambiguity, vagueness and synonymy) and provides a higher probability of a correct semantic matching (Freitas and Curry 2014). This first alignment is called a *semantic pivot* and typically maps to a *named entity* (rigid designator) in the query (which most likely maps to a constant in the database) and can be determined using linguistic-based heuristics (e.g. Part-of-Speech, corpus-based specificity/entropy measures such as IDF (Freitas and Curry 2014)) (Figure 2).

The selection of a semantic pivot provides a drastic *reduction of the symbolic matching space* (and the associated

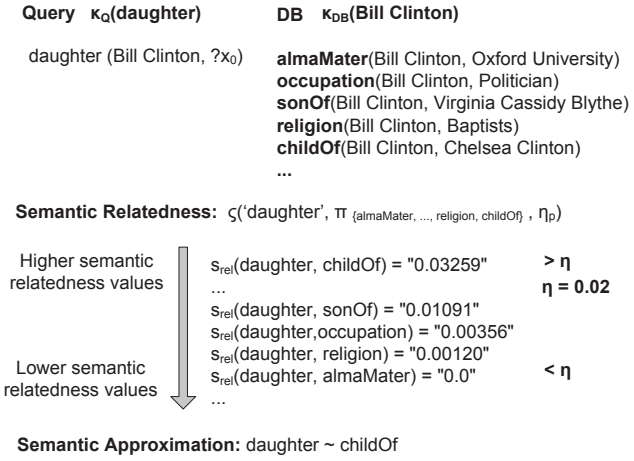

Figure 2: Distributional semantics approximation after the semantic pivot selection.

semantic entropy) for the semantic approximation mechanism of the following alignments, since they are restricted by the syntactic constraints of the semantic pivot.

D-DBMS Architecture

From an architectural perspective, a *database management system* (DBMS) can be enriched with a *separate distributional semantics approximation layer*. A high-level architecture diagram for the distributional DBMS (D-DBMS) is depicted in Figure 3. The *distributional semantic model component* builds the semantic model from the *reference corpora*, which can coincide with the target dataset. Different distributional models have different approximation properties, allowing broader or narrower semantic approximations, depending on the configuration of their parameters. Additionally, different reference corpora can be used according to the domain of discourse covered in the database. The semantic relatedness measure can be computed in two scenarios: (i) *dynamically:* where the semantic approximation operator ζ calculates the semantic relatedness between a query term and database elements, (ii) using a *distributional index:* where the distributional vectors of the database elements are represented in an inverted index (Freitas and Curry 2014). The semantic approximation layer can be implemented as an external, distinct layer from the database component, not requiring direct adaptation of the database internals.

Application Patterns

The convergence between distributional semantics and databases is a recent development which has been used in different application scenarios:

Schema-agnostic Queries: In this category the $\mathcal{RC}$ is unstructured and it is distinct from the $\mathcal{DB}$. (Freitas and Curry

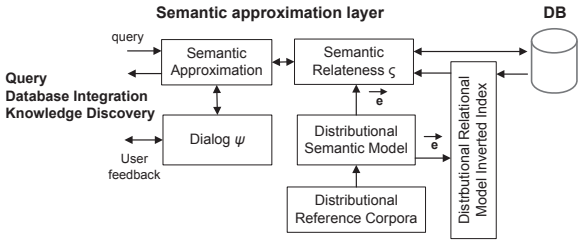

Figure 3: A distributional semantics layer to complement structured data semantics.

2014) define a DRM ($\tau - Space$) for supporting schema-agnostic queries over large-schema databases, where users are not aware of the representation of the data. The DSM used is Explicit Semantic Analysis (ESA). The approach was evaluated using natural language queries (QALD 2011 test collection) over DBpedia, a large-schema RDF graph dataset containing 45,767 properties, 9,434,677 instances and over 200,000 classes achieving *avg. recall=0.81*, *mean avg. precision=0.62* and *mean reciprocal rank=0.49*, with interactive-level query execution time (most queries are answered in less than 2s).

Selective Commonsense Reasoning over Incomplete KBs: (Freitas et al. 2014) uses a DRM to support selective reasoning over incomplete commonsense $\mathcal{KB}s$. Distributional semantics is used to select the facts which are semantically relevant under a specific (abductive-style) reasoning context, allowing the scoping of the reasoning context and also coping with incomplete knowledge of the commonsense KBs.

Approximative Logic Programming over Incomplete KBs: (da Silva and Freitas 2014) used a DRM to support approximate reasoning in logic programs, defining predicate substitutions to support the application of schema-agnostic queries and rules. The approach supports a runtime predicate integration within logic programs.

Knowledge Discovery: In this category, the structured $\mathcal{DB}$ is used as a distributional reference corpora (where $\mathcal{RC} = \mathcal{DB}$). 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, Handschuh, and Decker 2011) describe an approach for using web data as a bottom-up phenomena, capturing meaning that is not associated with explicit semantic descriptions. They apply it to entity consolidation in the life sciences domain. (Speer, Havasi, and Lieberman 2008) 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, Schvaneveldt, and Rindflesch 2009) introduced PSI, a predication-based semantic indexing for biomedical data. PSI was used for similarity-based retrieval and detection of implicit associations.

Conclusion

Preliminary results for *Distributional Relational Models* (DRMs) have been encouraging, showing the effectiveness of distributional semantics as a semantic model complementary to structured data semantics. Distributional semantics have been used to support semantic approximations for schema-agnostic queries, coping with knowledge bases' incompleteness for reasoning, and for knowledge discovery (entity consolidation, link discovery) in structured data.

Acknowledgments: This publication was supported in part by Science Foundation Ireland (SFI) (Grant Number SFI/12/RC/2289) and by the Irish Research Council.

References

Cohen, T.; Schvaneveldt, R. W.; and Rindflesch, T. C. 2009. Predication-based semantic indexing: Permutations as a means to encode predications in semantic space. In *T. AMIA Annu Symp Proc.*, 114–118.

da Silva, J. C. P., and Freitas, A. 2014. Towards an approximative ontology-agnostic approach for logic programs. In *Proc. Intl. Symposium on Foundations of Information and Knowledge Systems*.

Freitas, A., and Curry, E. 2014. Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. In *Proc. Intl. Conference on Intelligent User Interfaces*.

Freitas, A.; Silva, J. C. P. D.; Curry, E.; and Buitelaar, P. 2014. A distributional semantics approach for selective reasoning on commonsense graph knowledge bases. In *Proc. Intl. Conference on Applications of Natural Language to Information Systems*.

Freitas, A.; Silva, J. C. P. D.; and Curry, E. 2014. On the semantic mapping of schema-agnostic queries: A preliminary study. In *Proc of. NLIWoD at ISWC*.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the International Joint Conference on Artifical intelligence*, 1606–1611.

Harris, Z. 1954. Distributional structure. *Word 10 (23)* 146–162.

Novacek, V.; Handschuh, S.; and Decker, S. 2011. Getting the meaning right: A complementary distributional layer for the web semantics. In *Proc. of the Intl. Semantic Web Conference*, 504–519.

Speer, R.; Havasi, C.; and Lieberman, H. 2008. Analogyspace: Reducing the dimensionality of common sense knowledge. In *Proc. of the Intl. Conf. on Artificial Intelligence*, 548–553.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research* 141–188.