# Distributional Relational Networks

**André Freitas**[1], **João C. P. da Silva**[1,2], **Seán O'Riain**[1], **Edward Curry**[1]
[1]Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
[2]Computer Science Department
Federal University of Rio de Janeiro

### Abstract

This work introduces Distributional Relational Networks (DRNs), a Knowledge Representation (KR) framework which focuses on allowing semantic approximations over large-scale and heterogeneous knowledge bases. The proposed model uses the distributional semantics information embedded in large text/data corpora to provide a comprehensive and principled solution for semantic approximation. DRNs can be applied to open domain knowledge bases and can be used as a KR model for commonsense reasoning. Experimental results show the suitability of DRNs as a semantically flexible KR framework.

## Introduction

Relational and logical models provide an expressive system for representing concepts, objects, their attributes and associations. In addition to the representation of conceptual abstractions, these models provide formalized definitions for operations such as querying and logical inference.

Despite their ability to provide an expressive and principled representation, existing models have practical limitations for delivering a knowledge representation (KR) framework which is able to cope with conceptual model heterogeneity, inconsistency, contextual complexity, vagueness and ambiguity. These requirements become present when the models start to move outside the controlled environment of domain specific and manually created models, moving in the direction of large-scale open domain models.

To provide additional flexibility and cope with variations in conceptualizations, most KR approaches today depend on the explicit addition of statements and rules to the KB. These statements and rules can automatically materialize new statements in the KB under alternative conceptualizations using a deductive reasoning model. Under this perspective the ability to provide a KB which is able to cope with all possible conceptual models for all possible KB users depends on the following assumptions: (i) on the ability of the KB designer to anticipate all possible conceptual models, (ii) on the ability to generate a large set of statements and rules to cope with all possible conceptual models, (iii) while keeping consistency and scalability. This KR model does not scale to large KBs.

The lack of properties such as a principled mechanism of semantic approximation which is not dependent on the construction of a large-scale consistent commonsense KB, and the ability to scale to large volume knowledge bases proved in practice to be a strong limitation of most of the existing KR frameworks.

More recently, *distributional semantics* [1] have emerged from the empirically supported evidence that models derived from statistical co-occurrence patterns on large collections of text can provide simplified but comprehensive semantic models. Distributional semantic models can be automatically built from large corpora, not requiring manual construction effort on the creation of the semantic model.

With the availability of large volumes of text on the Web, comprehensive distributional semantic models can be built. Distributional semantic models also provide a *quantitative* perspective on semantics, which can be used in the process of semantic matching and approximation. Additionally, distributional semantic models are associated with vector space models (VSMs), where existing dimensional reduction approaches or inverted list indexes can provide the scalability for the instantiation of large-scale distributional models.

While theoretical and applied aspects of distributional semantic models (DSMs) have been investigated in the Computational Linguistics and Information Retrieval circles, the interaction between DSMs and KR models is not yet fully explored.

This work analyzes the complementary aspects between *distributional semantics* and *relational/logic-based KR models*. *Distributional Relational Networks* (DRNs) are introduced as a KR framework which unifies these two perspectives, where the relational/graph structure provides the fine-grained semantics and it is complemented by the distributional model, which works as a *large-scale coarse-grained associational structure*. DRNs provide a principled and built-in mechanism to include semantic approximation in the process of querying and reasoning over KBs, allowing the embedding and usage of large-scale unstructured and structured commonsense information into the querying and reasoning process.

The property of embedding large-scale conmmonsense information into KBs allows a common semantic integration ground between KBs with different conceptual models. The proposed framework can be applied to any KR model which can be formulated as a labelled graph structure, giving generality to its application over different KR models.

## Motivational Scenario

Open domain/commonsense reasoning is an important part of AI application scenarios such as question-answering systems and it demands approaches which can cope with the intrinsic contextual semantic complexity and KB scale. Every knowledge or information artifact (from unstructured text to structured knowledge bases) maps to an implicit or explicit set of user intents and semantic context patterns. The multiplicity of contexts where open domain and commonsense knowledge bases can be used, defines an intrinsic semantic heterogeneity scenario for these uses.

Different levels of conceptual abstraction or lexical expressions in the representation of relations and entities are examples where a semantic gap can strongly impact the inference process. This section introduces the challenges which are the focus of this work by using a concrete example.

Consider that an user wants to ask the query *''Is the mother in law of Stanley Robinson's son an artist?''* to a given knowledge base Kb formed by the following set of facts and rules:

$$childOf(katehudson, goldiehawn).$$
$$childOf(chrisrobinson, stanleyrobinson).$$
$$spouse(katehudson, chrisrobinson).$$
$$isanActress(goldiehawn)$$
$$motherInLaw(A,B) \leftarrow spouse(B,C) \wedge childOf(C,A)$$

meaning that Kate Hudson is the child of Goldie Hawn, Chris Robinson is the child of Stanley Robinson, Kate Hudson is the spouse of Chris Robinson, Goldie Hawn is an actress and A is mother in law of B when the spouse of B is a child of A.

Suppose that the user is not aware of the terms and concepts inside Kb, while querying it: $?-sonOf(X, stanleyrobinson) \wedge motherInLaw(Y,X) \wedge isanArtist(Y)$.

The inference over Kb will not materialize the answer $X = chrisrobinson$ and $Y = goldiehawn$, because despite the statement and the rule describing the same subdomain, there is no precise vocabulary matching between the query and the Kb.

In order for the reasoning to work, the *semantic approximation* of the following terms would need to be established: 'sonOf' $\sim$ 'childOf', 'isanArtist' $\sim$ 'isanActress'. To close the *semantic/vocabulary gap* in a traditional deductive knowledge base it would be necessary to increase the size of the Kb to such an extent that it would contain all the facts and rules necessary to cope with any potential vocabulary difference. Together with the aggravation of the scalability problem, it would be necessary to provide a principled mechanism to build such a large scale and consistent set of facts and rules.

These are limitations of most of the existing KR approaches. To cope with semantic approximation, a KR framework should be able to address the following requirements:

1. *Ability to cope with lexical expression differences:* Concepts with strongly related meanings may have different lexical expressions. For example, the predicate 'is husband of' is a gender specific expression of the concept 'spouse'. Lexical variations can cross grammatical classes' boundaries: the closest term mapping to a verb in a query may be expressed as a noun in a $KB$.
2. *Ability to cope with abstraction level differences:* Differences in the core concept structures between the database representation and the concepts used in the query. For example an attribute named 'is an Actress' and another predicate/attribute 'is an Artist' express two different sets where the former set is contained in the second. In some cases the abstraction level expressed in the query can be different from the dataset.
3. *Ability to cope with compositional/structural differences:* Information may be organized in different KB structures. The attribute 'is an Actress' can be expressed as a unary attribute or can be expressed as the binary relation 'occupation' and an associated entity/value 'Actress'.
4. *Comprehensive commonsense KB:* The ability to semantically interpret and approximate information is largely dependent on the volume of commonsense knowledge available. The KR should have an associated comprehensive commonsense KB and should be able to use commonsense information in the query and reasoning process.
5. *Performance and Scalability:* The KR framework should allow approximations for query and reasoning to scale over large KBs.

## 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 [1]. A rephrasing of the *distributional hypothesis* states that words that occur in similar contexts tend to have similar/related meaning [1].

Distributional semantics focuses on the construction of a semantic representation of a word based on the statistical distribution of word co-occurrence in texts. The availability of high volume and comprehensive Web corpora brought distributional semantic models as a promising approach to build and represent meaning. Distributional semantic models are naturally represented by *Vector Space Models* (VSM), where the meaning of a word is represented by a *weighted concept vector*.

However, the proper use of the simplified model of meaning provided by distributional semantics implies understanding its characteristics and limitations. In distributional semantics, *differences of meaning* are mediated by *differences of distribution* in a reference corpora. As a consequence, distributional semantic models allow the *quantification* of the amount of association or difference in meaning between linguistic entities. This can be used to quantify the *semantic relatedness between words*. The intuition behind this approach is that two terms which are highly semantically related in a distributional model are likely to have a close (implicit) relation. Note that distributional semantic models can be specialized to exclude certain types of semantic relatedness such as antonyms or relations in a negation context. The computation of semantic relatedness between pairs of words is one instance in which the strength of distributional models and methods is empirically supported [2].

There are three core elements at the center of the Distributional Relational Network (DRN) model: (i) the use of *semantic relatedness measures* as a *principled semantic approximation operation* for *querying* and *reasoning* (Q&R) over conceptual and lexical KR models; (ii) the use of distributional semantics to build the semantic relatedness measures; (iii) the use of a compositional model for querying and reasoning over the relational structure.

## Semantic Relatedness

The concept of *semantic relatedness* is described [5] as a generalization of *semantic similarity*, where semantic similarity is associated with taxonomic relations between concepts (e.g. *car* and *airplane* share *vehicle* as a common taxonomic ancestor) and semantic relatedness covers a broader range of semantic relations (e.g. *car* and *driver*). Since differences in conceptual models can both cross *taxonomical* and *grammatical class* borders, the more generic concept of semantic relatedness is more suitable to the task of semantic approximation over these datasets.

Until recently, resources such as WordNet were used in the computation of semantic similarity and relatedness measures. The limitations of the representation present in WordNet include the lack of a rich representation of non-taxonomic relations (fundamental for the computation of semantic relatedness measures) and a limited number of modelled concepts. The availability of large amounts of unstructured text on the Web motivated the creation of semantic relatedness measures based on large text collections using distributional semantic models. These measures focus on addressing the limitations of resource-based approaches by trading structure for volume of commonsense knowledge [2].

Comparative evaluations between WordNet-based and distributional approaches for the computation of semantic relatedness measures have shown the strength of the distributional model, reaching a high correlation level with human assessments [2,4].

## Distributional Semantics Principles

In a distributional model, the observation of the reality is mediated by a *reference data corpus*, which captures a set of *observation views* of the reality under a *symbolic representation scheme*. The representation scheme is defined by a set of symbols which may be unstructured (e.g. as in natural language texts).

The relation between objects obeys a locality principle which can be related to the *spatio-temporal locality* (e.g. dependent on the distance of these objects on space and time) or to a *categorical locality*, where objects are similar in relation to a set of features. In the corpus, a co-occurrence context can be defined by different spatio-temporal locality criteria (in a natural language text a context can be a sentence, paragraph or document) (Figure 1).

Distributional semantic models can be represented as a vector space, where each dimension represents a context identifier or a co-occurring symbol(word) in the corpus. The distributional vector space supports the definition of a geometric interpretation for each symbol in relation to other symbols in the corpus, and provides a principled process for approximating two symbols (words), which consists in the calculation of a similarity measure between the interpretation vectors (e.g. cosine similarity).

A *co-occurrence context set* $c_n$ is defined by a set of co-occurring symbols in a context defined over the corpus. The *distributional interpretation* $[[s]]$ of a symbol s is defined by integrating all the co-occurrence contexts sets of the symbol and by defining a *membership degree* associated with each co-occurrence context set. The *membership degree* is a function of the co-occurrence frequency in each context in the corpus and defines a specificity measure by weighting out co-occurrence patterns which are shared among different contexts sets (e.g. symbol frequency/inverse context frequency).

Once a distributional space is built for a set of symbols $S$, new structured and unstructured data can be embedded in the space using the interpretation reference frame from another reference corpus. The *atomic context-level* is defined by the *compositional* (syntactic) structure of symbols and can be used to define a *relational structure* over symbols, which can be represented as relational vectors ($r$) in the distributional space.

## Distributional-Compositional Models

Distributional semantic models are complemented by compositional models which provide principled mechanisms to compose the meaning of multiple distributional interpretation vectors.

Clark & Pulman [6] provide a formal description of a compositional model of meaning, where distributional models are unified with a compositional theory of grammatical types. The approach focuses on the unification of the quantitative strength of distributional approaches with the compositionality provided by symbolic approaches. The final mathematical structure uses vectors to represent word meanings, and the tensor product to allow the composition of meaning and types. Coecke et al. [7] addresses some of the shortcomings present in the model of Clark & Pulman [6] proposing a generalized mathematical framework for a compositional distributional model of meaning.

Erk & Pado [9] introduce a structured vector space model which integrates syntax into the computation of word meaning in its syntactic context. Baroni & Lenci [8] propose a distributional semantic memory, a graph of weighted links between words which can be specialized to different corpus-based semantics tasks. Both works propose models that induce graphs describing the corpus syntactic relations.

While these compositional models propose the integration of distributional semantics and syntactic models, this work proposes the integration between distributional and relational models, approaching compositionality under the perspective of a semantic representation, exploring the connections with KR.

## Distributional Relational Networks (DRNs)

Distributional Relational Networks (DRN) merge *relational* and *distributional* representation models in order to allow semantic approximation over existing relational models. In the context of this work, a relational model is defined as a *labelled graph* where all symbols associated with nodes (entities) and edges (attributes/relations) have corresponding elements in a *reference data corpora* (unstructured or structured data collection used to build the distributional model). DRNs can be applied to different KR frameworks which can be mapped to the generic labelled graph representation, including semantic networks, logical KBs and relational/Entity-Attribute-Value(EAV) databases.

A DRN embeds the structure defined by relational models in a distributional vector space $c_n$. Every entity and relation has an associated weighted concept vector representation in the distributional concept space. The distributional associational information embedded in the distributional concept space is used to semantically complement the knowledge expressed in the relational model (Figure 1). The distributional information is then used to allow approximative querying/reasoning (Q&R) processes, differently from what happens in a relational model, where Q&R processes are constrained by strict syntactical definitions of entities and relations, and only exact matches allow the Q&R processes to continue.

Once a knowledge base Kb is embedded in a distributional space, all the symbols and its associations as well as the Q&R processes have an associated geometric representation. These processes have an associative nature, where the relational graph is navigated, guided by the *semantic relatedness* matching between the external user query or reasoning terms (i.e. the *users' semantic intent*). The semantic relatedness measure works as a *semantic heuristics*, guiding the Q&R process in the direction of the Q&R answer and reconciling the common intent of both conceptual models (the Kb and querying), independent of the vocabulary used to express it.

Another important characteristic of DRNs is that they are not committed to a particular relational model neither with a particular distributional model, allowing the combination of different models. The following subsections detail some of the elements of the DRN model.

### Relational Model

The relational model has a signature $\Sigma = (P, E)$ formed by a pair of finite set of symbols used to represent relations $p \in P$ between entities $e \in E$. We assume that both elements in $P$ and $E$ are represented using distributionally meaningful descriptors (symbols present in the reference corpus).

The signature is used, in conjunction with a set of operators to construct a knowledge base Kb. Each element in the signature $\Sigma_{Kb}$ is represented as a vector in a distributional space. The semantics of Kb is defined by the vectors in the distributional space used to represent the elements of Kb.

### Geometrical Model

The DRN space is named *T-Space* [3]. The *T-Space* is a distributional structured vector space model which allows the representation of the elements of a KB under a distributional semantic model.

The *T-Space coordinate system* is built from a document collection $\mathbb{C}$. The set $Term = \{k_1, \cdots, k_t\}$, of all terms available in $\mathbb{C}$ is used to define the basis $Term_{base} = \{\overrightarrow{\mathbf{k}}_1, \cdots, \overrightarrow{\mathbf{k}}_t\}$ of unit vectors that spans the *term vector space* $VS^{Term}$.

The set of all distributional concepts $Concept = \{c_1, \cdots, c_t\}$ are extracted from a reference corpus and each concept $c_i \in Concept$ is mapped to an identifier which represents the co-occurrence pattern in the corpus. Each identifier $c_i$ defines a set which tracks the context where a term $k_t$ occurred. This set is used to construct the basis $Concept_{base} = \{\overrightarrow{\mathbf{c}}_1, \cdots, \overrightarrow{\mathbf{c}}_t\}$ of vectors that spans the *distributional vector space* $VS^{dist}$ (Figure 1).

Thus, the set of contexts where a term occurs define the concept vectors associated with the term, which is a representation of its meaning on the reference corpus. Each concept vector is weighted according to the term distribution in the corpus, allowing the concept vector space coordinate basis to be defined in terms of a term vector space coordinate basis where each dimension maps to a word in the corpus. So, a vector $\overrightarrow{\mathbf{x}} \in VS^{dist}$ can be mapped to $VS^{Term}$ by the application of the following transformation:

$$\overrightarrow{\mathbf{x}} = \sum_{i=1}^{t} \alpha_i v_i^x \overrightarrow{\mathbf{k}}_i \tag{1}$$

where $v_i^x$ is the term co-occurrence pattern over a corpus and $\alpha_i$ is a second-order transformation tensor which is defined by the set of term vectors of distributional concepts.

### DRNs: Linking Relational and Geometrical Models

In order to obtain an approach that supports an approximative semantic Q&R model, we link the relational and geometrical models so that the geometrical model could enrich and ground the semantics of the relational model.

The first step is to build the *T-Space concept space* based on the reference corpus.

The second step is to translate the elements of the signature $\Sigma = (P, E)$ of a $KB$ to elements of $VS^{Term}$ and $VS^{dist}$. The vector representation of $P$, respectively, in $VS^{Term}$ and $VS^{dist}$ is defined by:

$$\overrightarrow{\mathbf{P}}_{VS^{Term}} = \{\overrightarrow{\mathbf{p}} : \overrightarrow{\mathbf{p}} = \sum_{i=1}^{t} w_i^p \overrightarrow{\mathbf{k}}_i, \text{for each } p \in P\} \tag{2}$$

$$\overrightarrow{\mathbf{P}}_{VS^{dist}} = \{\overrightarrow{\mathbf{p}} : \overrightarrow{\mathbf{p}} = \sum_{i=1}^{t} v_i^p \overrightarrow{\mathbf{c}}_i, \text{for each } p \in P\} \tag{3}$$

and the vector representation of $E$, respectively, in $VS^{Term}$ and $VS^{dist}$ is defined by:

$$\overrightarrow{\mathbf{E}}_{VS^{Term}} = \{\overrightarrow{\mathbf{e}} : \overrightarrow{\mathbf{e}} = \sum_{i=1}^{t} w_i^e \overrightarrow{\mathbf{k}}_i, \text{ for each } e \in E\} \quad (4)$$

$$\overrightarrow{\mathbf{E}}_{VS^{dist}} = \{\overrightarrow{\mathbf{e}} : \overrightarrow{\mathbf{e}} = \sum_{i=1}^{t} v_i^e \overrightarrow{\mathbf{c}}_i, \text{ for each } e \in E\} \quad (5)$$

where $w_i^e$ and $w_i^p$ are defined by co-occurrence weighting scheme[1] and $v_i^e$ and $v_i^p$ are defined by the weighting scheme over the distributional model.

The third step refers to the translation of Kb atoms into *T-Space* elements. As each relation and entity symbol has a vector representation, we can define the vector representation of a relational atom $r$ in the concept vector space by the following definition.

**Definition:** Let $\overrightarrow{\mathbf{p}}$, $\overrightarrow{\mathbf{e_1}}$ and $\overrightarrow{\mathbf{e_2}}$ be the vector representations, respectively, of $p, e_1$ and $e_2$. An atom vector representation (denoted by $\overrightarrow{\mathbf{r}}$) is defined by: $(\overrightarrow{\mathbf{p}} - \overrightarrow{\mathbf{e_1}})$ if $p(e_1)$; $(\overrightarrow{\mathbf{p}} - \overrightarrow{\mathbf{e_1}}, \overrightarrow{\mathbf{e_2}} - \overrightarrow{\mathbf{p}})$ if $p(e_1, e_2)$.

## Querying & Reasoning

The embedding of Kb in the distributional vector space allows the definition of a geometric interpretation for the Q&R processes. The proposed Q&R model uses the *cosine similarity* ([3]) as a semantic approximation and navigation operation in the *T-Space*. The distributional semantic relatedness measure can be used to establish an approximate semantic equivalence between two elements in the context of a given Q&R navigation step.

**Semantic Relatedness** A *semantic relatedness function* $sr : VS^{dist} \times VS^{dist} \to [0, 1]$ is defined as $sr(\overrightarrow{\mathbf{p_1}}, \overrightarrow{\mathbf{p_2}}) = \cos(\theta) = \overrightarrow{\mathbf{p_1}}.\overrightarrow{\mathbf{p_2}}$. A threshold $\eta \in [0, 1]$ could be used to establish the semantic relatedness between the two vectors: $sr(\overrightarrow{\mathbf{p_1}}, \overrightarrow{\mathbf{p_2}}) > \eta$.

**Querying & Reasoning: Matching and Navigation** The first element to be resolved in the ordered query, called *the semantic pivot*, normally is a symbol which represents an entity. The semantic pivot, as the more constraining element in the query, helps to reduce the search space since just the elements in Kb associated with the pivot at a given iteration are candidates for the semantic matching. Note that the query sequence is embedded in the vector space $VS^{dist}$, allowing to identify it with the following sequence of vectors $< \overrightarrow{\mathbf{q'}}_0, \overrightarrow{\mathbf{q'}}_1, \cdots, \overrightarrow{\mathbf{q'}}_n >$.

**Definition:** Given a query $q$, its entities and relations, denoted by $q_0, q_1, ..., q_n$ are ordered in a sequence $< q'_0, q'_1, \cdots, q'_n >$ using a heuristic measure of specificity $h_{specificity}$ from the most specific to the less specific, that is, $\forall i \in [0, n], h_{specificity}(q'_i) \geq h_{specificity}(q'_{i+1})$.

The goal behind this heuristic is to force the reasoning process to prioritize the hardest constraints in the query, which normally have the less semantic ambiguity [2].

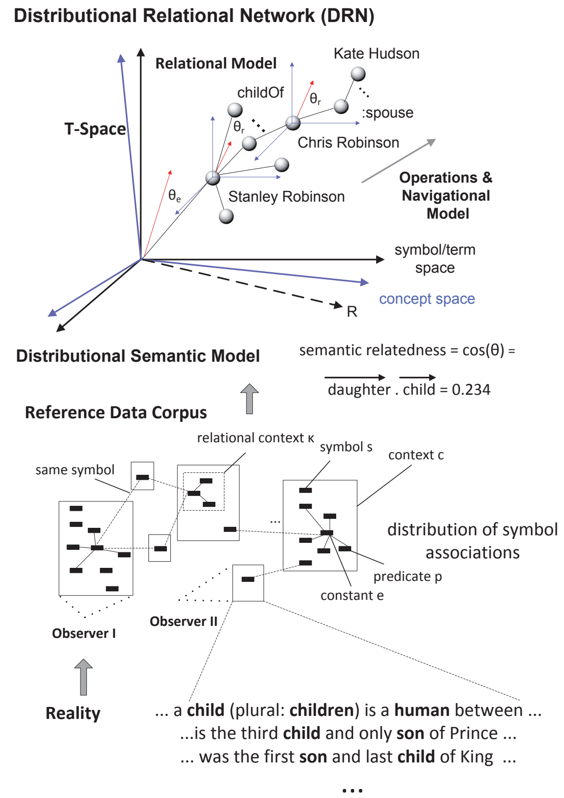


Figure 1: Depiction of a DRN construction and query workflow.

In the first iteration, $\overrightarrow{\mathbf{q'}}_0 \in VS^{dist}$, the vector representation of the pivot $q'_0$ can be resolved to a vector $\overrightarrow{\mathbf{e}}_0$ (Figure 1). The entity $e_0$ defines a vector subspace which can be explored by the next query term (which spans the relations associated with the entity $e_0$). The second query term $q'_1$ can be matched with one or more relations associated with $e_0$, for example $p_0$, considering that $sr(\overrightarrow{\mathbf{q'}}_1, \overrightarrow{\mathbf{p}}_0) \geq \eta$, where $\eta$ is a semantic relatedness threshold. The entities associated with $p_0$ (for example $e_1$) are used as new semantic pivots.

At each iteration of the (Q&R) process, a set of semantic pivots are defined and are used to navigate to other points in the $VS^{dist}$. This navigation corresponds to the reconciliation process between the semantic intent defined by the query and the semantic intent expressed in the KB. The reconciliation process can be defined as the sequence of vectors $< (\overrightarrow{\mathbf{q'}}_1 - \overrightarrow{\mathbf{p}}_1), (\overrightarrow{\mathbf{q'}}_2 - \overrightarrow{\mathbf{p}}_2), \cdots, (\overrightarrow{\mathbf{q'}}_n - \overrightarrow{\mathbf{p}}_n) >$. The proposed approximate Q&R process can also be represented geometrically as the vectors $< (\overrightarrow{\mathbf{e}}_0 - \overrightarrow{\mathbf{p}}_0), (\overrightarrow{\mathbf{p}}_0 - \overrightarrow{\mathbf{e}}_1), ..., (\overrightarrow{\mathbf{p}}_{n-1} - \overrightarrow{\mathbf{e}}_n) >$ over the T-Space, which represents the process of finding the answer in the DRN.

## Discussion

The quality of the semantic approximation in the Q&R process over DRNs is dependent on the quality of distributional models and on the intrinsic ambiguity of human lan-

[1] for example, the term-frequency/inverse document frequency(TF/IDF).

[2] in practice this specificity measure can be defined by a combination of grammatical classes weights and TF/IDF.

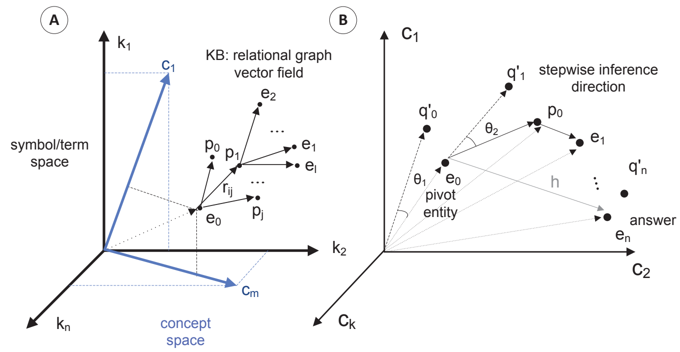


Figure 2: Vector representation for entities and relations.

guage. Despite the selectivity of the distributional model some Q&R processes may return spurious answers together with the relevant answers (as in information retrieval scenarios). Principled disambiguation and user dialogs operators can be defined at each semantic approximation step to increase its accuracy.

While pure relational models demand an a priori reconciliation of the symbols in a consistent conceptual model, distributional semantics allows the quantification of different usage contexts of a symbol. The distributional semantics representation model captures the superposition of different contexts. The disambiguation process can be performed under reasoning time, by either providing additional contextual information or by the interaction with an external agent in the model. The capture of superposition of different senses for a symbol, motivated attempts to bridge distributional semantics with models based on formalisms of Quantum Mechanics (Hilbert Spaces).

Relational graphs from different domains can be supported by different distributional models and different distributional reference corpora. Spaces with different distributional models can form patches in a more complex distributional manifold. Additionally, different distributional models can be used in parallel to support multiple interpretation of the elements embedded in the space.

An initial DRN model was implemented in [3], which proposes a structured vector space model (T-Space) targeting vocabulary-independent (schema-agnostic) and open domain natural language queries over heterogeneous Semantic Web data. The vector space is built using Explicit Semantic Analysis (ESA) as a distributional model and Wikipedia as a reference data corpus. The approach was evaluated using DBpedia, a heterogeneous graph database containing 45,767 predicates, 5,556,492 classes and 9,434,677 instances and the Question Answering over Linked Data test collection (50 complex natural language queries)[3], and achieved *avg. recall=0.491*, *mean avg. precision=0.482* and *mean reciprocal rank=0.516* [3].

More recent experiments done by the authors with an increased query set (102 natural language queries) for the same dataset showed achieved *avg. recall=0.81*, *mean avg. precision=0.62* and *mean reciprocal rank=0.49*.

The quality of the selectivity of distributional models such as ESA, as a semantic matching model was evaluated separately in [4], achieving avg. p@5=0.732. While ESA provides a comprehensive semantic model, where the semantic relatedness measure can be used as a ranking score [4], it does not provide absolute precision. Solutions to circumvent the limitation, which should be present in all distributional models, include the composition of distributional models which are more restrictive with more comprehensive approaches, and the application of user disambiguation operations. The quality of the semantic approximation provides some preliminary indication that DRNs can cope with requirements 1,2,3,4.

From the scalability and performance perspective, DRNs can be implemented as an *inverted index*, which can be segmented into parallel indexes, split by the entities in the graph. Experiments over DBpedia+YAGO datasets had an 8,530 ms average query execution time which supports requirement 5. Additionally, the approach provides a mechanism which demands no dataset adaptation effort, not requiring specific manual semantic enrichment.

**Acknowledgments.** This work has been funded by Science Foundation Ireland (Grant No. SFI/08/CE/I1380 (Lion-2)). João C. P. da Silva is a CNPq Fellow - Science without Borders (Brazil).

## References

[1]. P. D. Turney and P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research* 37, pp. 141-188, 2010.

[2]. E. Gabrilovich and S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in *Proc. Intl. Joint Conference On Artificial Intelligence*, 2007.

[3]. A. Freitas, E. Curry, J. G. Oliveira, S. O'Riain, A Distributional Structured Semantic Space for Querying RDF Graph Data. International Journal of Semantic Computing (IJSC), 2012.

[4]. A. Freitas, E. Curry and S. O'Riain, 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.

[5]. P. Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, In Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-95) 1995, pp. 448-453, 1995.

[6]. S. Clark and S. Pulman, Combining symbolic and distributional models of meaning, in *Proc. of the AAAI Spring Symposium on Quantum Interaction*, Stanford, CA, pp. 52-55, 2007.

[7]. B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical Foundations for a Compositional Distributional Model of Meaning, volume 36. *Linguistic Analysis* (Lambek Festschrift), 2010.

[8]. M. Baroni, A. Lenci, Distributional memory: A general framework for corpus-based semantics. Comput. Linguist. 36(4), 673–721, 2010.

[9]. K. Erk, S. Pado, A structured vector space model for word meaning in context, in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[3]http://www.sc.cit-ec.uni-bielefeld.de/qald-1