A Distributional Semantics Approach for Selective Reasoning on Commonsense Graph Knowledge Bases

André Freitas[1], João C. P. da Silva[1,2], Edward Curry[1], Paul Buitelaar[1,3]

[1]Insight Centre for Data Analytics, National University of Ireland, Galway
[2]Computer Science Department, Federal University of Rio de Janeiro
[3]College of Graduate Studies, University of South Africa

Abstract. Tasks such as question answering and semantic search are dependent on the ability of querying & reasoning over large-scale commonsense knowledge bases (KBs). However, dealing with commonsense data demands coping with problems such as the increase in schema complexity, semantic inconsistency, incompleteness and scalability. This paper proposes a selective graph navigation mechanism based on a distributional relational semantic model which can be applied to querying & reasoning over heterogeneous knowledge bases (KBs). The approach can be used for approximative reasoning, querying and associational knowledge discovery. In this paper we focus on commonsense reasoning as the main motivational scenario for the approach. The approach focuses on addressing the following problems: (i) providing a semantic selection mechanism for facts which are relevant and meaningful in a specific reasoning & querying context and (ii) allowing coping with information incompleteness in large KBs. The approach is evaluated using ConceptNet as a commonsense KB, and achieved *high selectivity*, *high scalability* and *high accuracy in the selection of meaningful navigational paths*. Distributional semantics is also used as a principled mechanism to cope with information incompleteness.

1 Introduction

Building intelligent applications and addressing simple computational semantic tasks demand coping with large-scale commonsense Knowledge Bases (KBs). Querying and reasoning (Q&R) over large commonsense KBs are fundamental operations for tasks such as Question Answering, Semantic Search and Knowledge Discovery. However, in an open domain scenario, the scale of KBs and the number of direct and indirect associations between elements in the KB can make Q&R grow unmanageable. To the complexity of querying and reasoning over such large-scale KBs, it is possible to add the barriers involved in building KBs with the necessary consistency and completeness requirements.

With the evolution of open data, better information extraction frameworks and crowdsourcing tools, large-scale structured KBs are becoming more available. This data can be used to provide commonsense knowledge for semantic applications. However, querying and reasoning over this data demands approaches which are able to cope with large-scale, semantically heterogeneous and incomplete KBs.

As a motivational scenario, suppose we have a KB with the following fact: *'John Smith is an engineer'* and suppose the query *'Does John Smith have a degree?'* is issued

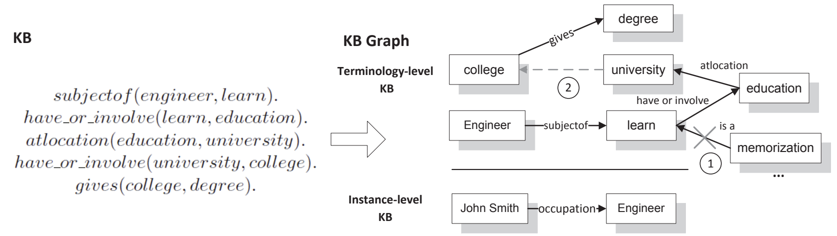

Fig. 1: (1) Selection of meaningful paths, (2) Coping with information incompleteness.

over the KB. A complete KB would have the rule *'Every engineer has a degree'*, which would materialize *'John Smith has a degree'*. For large-scale and open domain commonsense reasoning scenarios, model completeness and full materialization cannot be assumed. In this case the information can be embedded in other facts in the KB (Figure 1). The example sequence of relations between *engineer* and *degree* defines a path in a large-scale graph of relations between predicates, which is depicted in Figure 1.

In a large-scale KB, full reasoning can become unfeasible. A commonsense KB would contain vast amounts of facts and a complete inference over the entire KB would not scale to its size. Furthermore, while the example path is a meaningful sequence of associations for answering the example query, there is a large number of paths which are not meaningful under a specific query context. In Figure 1(1), for example, the reasoning path which goes through (1) is not related to the goal of the query (the relation between *engineer* and *degree*) and should be eliminated. Ideally a query and reasoning mechanism should be able to filter out facts and rules which are unrelated to the Q&R context. The ability to select the minimum set of facts which should be applied in order to answer a specific user information need is a fundamental element for enabling reasoning capabilities for large-scale commonsense knowledge bases.

Additionally, since information completeness of the KBs cannot be guaranteed, one missing fact in the KB would be sufficient to block the reasoning process. In Figure 1(2) the lack of a fact connecting university and college eliminates the possibility of answering the query. Ideally Q&R mechanisms should be able to cope with some level of KB incompleteness, approximating and filling the gaps in the KBs.

This work proposes a *selective reasoning approach* which uses a *hybrid distributional-relational semantic model* to address the problems previously described. Distributional semantic models (DSMs) use statistical co-occurrence patterns, automatically extracted from large unstructured text corpora, to support the creation of comprehensive quantitative semantic models. In this work, DSMs are used as complementary semantic layer to the relational model, which supports coping with semantic approximation and incompleteness. The proposed approach focuses on the following contributions:

- provision of a selective Q&R approach using a distributional semantics heuristics, which reduces the search space for large-scale KBs at the same time it maximizes paths which are more meaningful for a given reasoning context;

- definition of a Q&R model which copes with the information incompleteness present at the KB, using the distributional model to support semantic approximations, which can fill the lack of information in the KB during the reasoning process;

This work is organized as follows: section 2 provides an introduction on distributional semantics; section 3 describes the τ-Space distributional-relational semantic model which is used for the selection reasoning mechanism; section 4 describes the selective reasoning mechanism (*distributional navigational algorithm*); section 5 provides an evaluation of the approach using Explicit Semantic Analysis (ESA) as a distributional semantic model and ConceptNet [11] as KB; section 6 describes related work and finally, section 7 provides conclusions and future work.

2 Distributional Semantics

In this work *distributional semantics* supports the definition of an *approximative semantic navigational approach* in a knowledge base, where the graph concepts and relations are mapped to vectors in a *distributional vector space*.

Distributional semantics is defined upon the assumption that the context surrounding a given word in a text provides important information about its meaning [12]. It focuses on the construction of a semantic model for a word based on the statistical distribution of co-located words in texts. These semantic models are naturally represented by Vector Space Models (VSMs), where the meaning of a word can be defined by a weighted vector, which represents the association pattern of co-occurring words in a corpus.

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. Additionally, its natural association with VSMs, which are supported by dimensional reduction approaches or data structures such as inverted list indexes can provide a scalability benefit for the instantiation of these models.

The computation of *semantic relatedness measure* between words is one instance in which the strength of distributional models and methods is empirically supported ([3];[2]). The computation of the *semantic relatedness measure* is at the center of this work and it is used as a *semantic heuristics* to navigate in the KB graph, *where the distributional knowledge extracted from unstructured text is used as a general-purpose large-scale commonsense KB, which complements the knowledge present at the relational KB*.

3 τ-Space

The τ-*Space* [1] is a *distributional structured vector space model* which allows the representation of the elements of a graph KB under the grounding of a distributional semantic model. This work improves the formalisation on the definition of the τ-*Space*.

τ-*Space* is built from a *reference corpus* $RC = (Term, Context)$ formed by a set of terms $Term = \{k_1, \cdots, k_t\}$ and a set of context windows $Context = \{c_1, \cdots, c_t\}$.

The set $Term$ is used to define the basis $Term_{basis} = \{\overrightarrow{\mathbf{k}}_1, \cdots, \overrightarrow{\mathbf{k}}_t\}$ of unit vectors that spans the *term vector space* VS^{term}.

A context window c_j is represented in VS^{term} as:

$$\overrightarrow{\mathbf{c}}_j = \sum_{i=1}^{t} v_{i,j} \overrightarrow{\mathbf{k}}_i \qquad (1)$$

where $v_{i,j}$ is 1 if term k_i appears in context window c_j and 0 otherwise.

Analogously, the set of context windows $Context$ is used to define the basis $Context_{basis} = \{\overrightarrow{\mathbf{c}}_1, \cdots, \overrightarrow{\mathbf{c}}_t\}$ of vectors that spans the *distributional vector space* VS^{dist}. A given term x is represented in VS^{dist} as:

$$\overrightarrow{\mathbf{x}} = \sum_{j=1}^{t} w_j \overrightarrow{\mathbf{c}}_j \qquad (2)$$

such that

$$w_j = tf_j \times idf = \frac{freq_j}{count(c_j)} \times \log \frac{N}{n_{c_j}} \qquad (3)$$

where w_j is the product of the normalized term frequency tf_j (the ratio between the frequency of term x in the context window c_j and the number of terms inside c_j) and the inverse document frequency idf for the term x (the logarithm of the ratio of the total number of N context windows in the reference corpus RC and the number n_{c_j} of context containing the term x).

Thus, the set of context windows where a term occurs define the concept vectors associated with the term, which is a representation of its meaning on the reference corpus.

4 Embedding the Commonsense KB into the τ-Space

We consider that a commonsense knowledge base KB is formed by a set of *concepts* $\{v_1, \cdots, v_n\}$ and a set of *relations* $\{r_1, \cdots, r_m\}$ between these concepts, both represented as words or short phrases in natural language. Formally, a commonsense knowledge base KB is defined by a *labeled digraph* $G^{label}_{KB} = (V, R, E)$, where $V = \{v_1, \cdots, v_n\}$ is a set of nodes, $R = \{r_1, \cdots, r_m\}$ is a set of relations and E is a set of directed edges (v_i, v_j) labeled with relation $r \in R$ and denoted by (v_i, r, v_j). Alternatively, we can simplify the representation of the KB ignoring their relation labels: Let KB be commonsense knowledge base and $G^{label}_{KB} = (V, R, E)$ be its labeled digraph representation. A simplified representation of KB is defined by a *digraph* $G_{KB} = (V', E')$, where $V' = V$ and $E' = \{(v_i, v_j) : (v_i, r, v_j) \in E\}$. Given the (labeled) graph representation of KB, we have to embed it into the τ-Space. To do that we have to translate the nodes and edges of the graph representation of KB into a vector representation in VS^{dist}. The vector representation of $G^{label}_{KB} = (V, R, E)$ in VS^{dist} is $\overrightarrow{\mathbf{G}}^{label}_{KB_{dist}} = (\overrightarrow{\mathbf{V}}_{dist}, \overrightarrow{\mathbf{R}}_{dist}, \overrightarrow{\mathbf{E}}_{dist})$ such that:

$$\overrightarrow{\mathbf{V}}_{dist} = \{\overrightarrow{\mathbf{v}} : \overrightarrow{\mathbf{v}} = \sum_{i=1}^{t} u_i^v \overrightarrow{\mathbf{c}}_i, \text{ for each } v \in V\} \tag{4}$$

$$\overrightarrow{\mathbf{R}}_{dist} = \{\overrightarrow{\mathbf{r}} : \overrightarrow{\mathbf{r}} = \sum_{i=1}^{t} u_i^r \overrightarrow{\mathbf{c}}_i, \text{ for each } r \in R\} \tag{5}$$

$$\overrightarrow{\mathbf{E}}_{dist} = \{(\overrightarrow{\mathbf{r}} - \overrightarrow{\mathbf{v_i}}, \overrightarrow{\mathbf{v_j}} - \overrightarrow{\mathbf{r}}) : \text{for each } (v_i, r, v_j) \in E\} \tag{6}$$

u_i^v and u_i^r are defined by the weighting scheme over the distributional model[1].

5 Distributional Navigation Algorithm

Once the KB is embedded into the τ-Space, the next step is to define the navigational process in this space that corresponds to a selective reasoning process in the KB. The navigational process is based on the semantic relatedness function defined as: $sr : VS^{dist} \times VS^{dist} \to [0, 1]$ is defined as:

$$sr(\overrightarrow{\mathbf{p_1}}, \overrightarrow{\mathbf{p_2}}) = \cos(\theta) = \overrightarrow{\mathbf{p_1}}.\overrightarrow{\mathbf{p_2}}$$

A threshold $\eta \in [0, 1]$ can be used to establish the desired semantic relatedness between two vectors: $sr(\overrightarrow{\mathbf{p_1}}, \overrightarrow{\mathbf{p_2}}) > \eta$.

The information provided by the semantic relatedness function sr is used to identify elements in the KB with a similar meaning from the reference corpus perspective. The threshold was calculated following the semantic differential approach proposed in [2]. Multiword phrases are handled by calculating the centroid between the concept vectors defined by each word.

Algorithm 1 is the Distributional Navigation Algorithm (DNA) which is used to find, given two semantically related terms $source$ and $target$ wrt a threshold η, all paths from $source$ to $target$, with length l, formed by concepts semantically related to $target$ wrt η.

The $source$ term is the first element in all paths (*line 1*). From the set of paths to be explored (*ExplorePaths*), the DNA selects a path (*line 5*) and expands it with all neighbors of the last term in the selected path that are semantically related wrt threshold η and that does not appear in that path (*line 7-8*). The stop condition is $sr(target, target) = 1$ (*line 10-11*) or when the maximum path length is reached.

The paths $p =< t_0, t_1, \cdots, t_l >$ (where $t_0 = source$ and $t_l = target$) found by DNA are ranked (*line 14*) according to the following formula:

$$rank(p) = \sum_{i=0}^{l} sr(\overrightarrow{\mathbf{t_i}}, \overrightarrow{\mathbf{target}}) \tag{7}$$

Algorithm 1 can be modified to use a heuristic that allows to expand only the paths for which the semantic relatedness between all the nodes in the path and the target term

[1] Reflecting the word co-occurrence pattern in the reference corpus

Algorithm 1 Distributional Navigation Algorithm

```
INPUT

  – threshold: $\eta$
  – pair of terms $(source, target)$ such that $sr(\overrightarrow{\mathbf{source}}, \overrightarrow{\mathbf{target}}) > \eta$
  – path length: $l$

OUTPUT
$RankedPaths$: a set of ranked score paths $< (t_0, \cdots, t_l), score >$ such that $t_0 = source$ and $t_l = target$

1:  $t_0 \leftarrow source$
2:  $Paths \leftarrow \emptyset$
3:  $ExplorePaths \leftarrow [(< t_0 >, sr(\overrightarrow{\mathbf{t_0}}, \overrightarrow{\mathbf{target}}))]$
4:  while $ExplorePaths \neq \emptyset$ do
5:      remove $(< t_0, \cdots, t_k >, sr(\overrightarrow{\mathbf{t_k}}, \overrightarrow{\mathbf{target}}))$ from $ExploredPaths$
6:      if $k < l - 1$ then
7:          for all $(n \in neighbors(t_k) : sr(\overrightarrow{\mathbf{n}}, \overrightarrow{\mathbf{target}}) > \eta$ and $n \notin \{t_0, \cdots, t_k\})$ do
8:              append $(< t_0, \cdots, t_k, n >, sr(\overrightarrow{\mathbf{n}}, \overrightarrow{\mathbf{target}}))$ to $ExplorePaths$
9:          end for
10:     else if $k = l - 1$ then
11:         append $(< t_0, \cdots, t_k, target >, 1)$ to $Paths$
12:     end if
13: end while
14: $RankedPaths \leftarrow sort(Paths)$
15: return $RankedPaths$
```

increases along the path. The differential in the semantic relatedness for two consecutive iterations is defined as $\Delta_{target}(t_1, t_2) = sr(\overrightarrow{\mathbf{t_2}}, \overrightarrow{\mathbf{target}}) - sr(\overrightarrow{\mathbf{t_1}}, \overrightarrow{\mathbf{target}})$, for terms t_1, t_2 and $target$. This heuristic is implemented by including an extra test in the line 7 condition, i.e., $\Delta_{target}(t_k, n) > 0$.

6 Evaluation

6.1 Setup

In order to evaluate the proposed approach, the *τ-Space* was built using the *Explicit Semantic Analysis* (ESA) as the distributional model. ESA is built over Wikipedia using the Wikipedia articles as *context co-occurrence windows* and TF/IDF as a weighting scheme.

ConceptNet[11] was selected as the commonsense knowledge base. *ConceptNet* is a semantic network represented as a labeled digraph $G^{label}_{ConceptNet}$ formed by a set of nodes representing concepts and a set of labeled edges representing relations between concepts. ConceptNet is built by using a combination of approaches, including open information extraction tools, crowd-sourced user input and open structured data. Concepts and relations are presented in the form of words or short natural language phrases. The bulk of the semantic network represents relations between predicate-level words or expressions. Different word senses are not differentiated. Two types of relations can be found: (i) recurrent relations based on a lightweight ontology used by ConceptNet (e.g. *partOf*) and (ii) natural language expressions entered by users and open information extraction tools. These characteristics make ConceptNet a heterogeneous commonsense knowledge base. For the experiment, all concepts and relations that were not in English terms were removed. The total number of triples used on the evaluation was 4,797,719.

The distribution of the number of clauses per relation type is as follows: $= 1$ **(45,311)**, $1 < x < 10$ **(11,804)**, $10 \leq x < 20$ **(906)**, $20 \leq x < 500$ **(790)**, ≥ 500 **(50)**.

A test collection consisting of 45 (*source*, *target*) word pairs were manually selected using pairs of words which are semantically related under the context of the Question Answering over Linked Data challenge (QALD 2011/2012)[2]. Each pair establishes a correspondence between question terms and dataset terms (e.g. 'What is the *highest* mountain?' where *highest* maps to the *elevation* predicate in the dataset). 51 pairs were generated in total.

For each word pair (a, b), the navigational algorithm 1 was used to find all paths with lengths 2, 3 and 4 above a fix threshold $\eta = 0.05$, taking a as source and b as target and vice-versa, accounting for a total of 102 word pairs. All experimental data is available online[3].

6.2 Reasoning Selectivity

The first set of experiments focuses on the measurement of the selectivity of the approach, i.e. the ability to select paths which are related and meaningful to the reasoning context. Table 1 shows the average *selectivity*, which is defined as the ratio between the *number of paths selected using the reasoning algorithm* 1 by the *total number of paths* for each path length. The total number of paths was determined by running a depth-first search (DFS) algorithm.

For the size of ConceptNet, paths with length 2 return an average of 5 paths per word pair. For this distance most of the returned paths tend to be strongly related to the word pairs and the selectivity ratio tend to be naturally lower. For paths with length 3 and 4 the algorithm showed a very high selectivity ratio (0.153 and 0.0192 respectively). The exponential decrease in the selectivity ratio shows the scalability of the algorithm with regard to selectivity. Table 1 shows the average selectivity for DNA. The variation of DNA with the Δ criteria, compared to DNA, provides a further selectivity improvement (ϕ = (# of spurious paths returned by DNA / # of spurious paths returned by DNA + Δ)) $\phi(length2)$ = 1, $\phi(length3)$ = 0.49, $\phi(length4)$ = 0.20.

Table 1: Selectivity

Path Length	**Average Selectivity Agorithm 1**	**% Pairs of Words Resolved**	**Path Acuracy**
2	0,602	0,618	0,958
3	0,153	0,726	0,828
4	0,019	0,794	0,736

6.3 Semantic Relevance

The second set of experiments focuses on the determination of the *semantic relevance of the returned nodes*, which measures the expected property of the distributional semantic

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

[3] http://bit.ly/1p3PmHr

relatedness measure to serve as a heuristic measure for the selection of meaningful paths.

A gold standard was generated by two human annotators which determined the set of paths which are *meaningful* for the pairs of words using the following criteria: (i) all entities in the path are highly semantically related to both the source and target nodes and (ii) the entities are not very specific (unnecessary presence of instances, e.g. *new york*) or very generic (e.g. *place*) for a word-pair context. Only senses related to both source and target are considered meaningful.

The accuracy of the algorithm for different path lengths can be found in Table 1. The *high accuracy* reflects the effectiveness of the distributional semantic relatedness measure in the selection of meaningful paths. A systematic analysis of the returned paths shows that the decrease in the accuracy with the increase on path size can be explained by the higher probability on the inclusion of instances and classes with high abstraction levels in the paths.

From the paths classified as not related, 47% contained entities which are too specific, 15.5% too generic and 49.5% were unrelated under the specific reasoning context. This analysis provides the directions for future improvements of the approach (inclusion of filters based on specificity levels).

6.4 Addressing information incompleteness

This experiment measures the suitability of the distributional semantic relatedness measure to cope with KB incompleteness (gaps in the KB). 39 $< source, target >$ entities which had paths with length 2 were selected from the original test collection. These pairs were submitted as queries over the ConceptNet KB indexed on the VS^{dist} and were ranked by the semantic relatedness measure. This process is different from the distributional navigational algorithm, which uses the relation constraint in the selection of the neighbouring entities. The distributional semantic search mechanism is equivalent to the computation of the semantic relatedness between the query ($source\ target$) and all entities (nodes) in the KB. The threshold criteria take the top 36 elements returned.

Two measures were collected. *Incompleteness precision* measures the quality of the entities returned by the semantic search over the KB and it is given by *incompleteness precision = # of strongly related entities / # of retrieved entities*. The determination of the *strongly related entities* was done using the same methodology described in the classification of the semantic relevance. In the evaluation, results which were not highly semantically related to both source and target and were too specific or too generic were considered incorrect results. The **avg. incompleteness precision value of 0.568** shows that the ESA-based distributional semantic search provides a feasible mechanism to cope with KB incompleteness, suggesting the discovery of highly related entities in the KB in the reasoning context. There is space for improvement by the specialization of the distributional model to support better word sense disambiguation and compositionality mechanisms.

The *incompleteness coefficient* provides an estimation of the incompleteness of the KB addressed by the distributional semantics approach and it is determined by *incompleteness coefficient = # of retrieved ConceptNet entities with an explicit association / #*

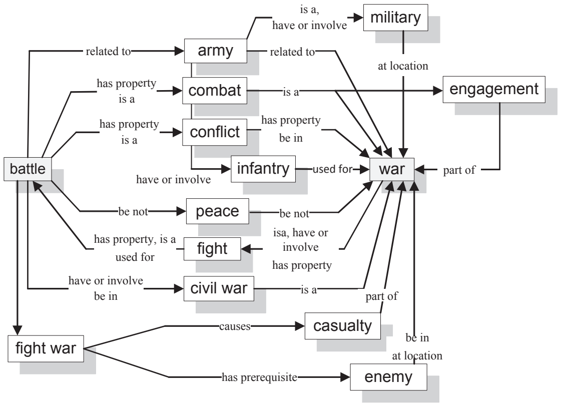

Fig. 2: Contextual (selected) paths between battle and war.

of strongly related retrieved entities. The **average incompleteness value of 0.039** gives an indication of the level of incompleteness that commonsense KBs can have. The *avg. # of strongly related entities* returned per query is 19.21.

An example of the set of new entities suggested by the distributional semantic relatedness for the pair $< mayor, city >$ are: **council, municipality, downtown, ward, incumbent, borough, reelected, metropolitan, city, elect, candidate, politician, democratic**.

The evaluation shows that distributional semantics can provide a principled mechanism to cope with KB incompleteness, returning highly related KB entities (and associated facts) which can be used in the reasoning process. The level of incompleteness of an example commonsense KB was analyzed and found to be high, confirming the relevance of this problem under the context of reasoning over commonsense KBs.

7 Analysis of the Algorithm Behavior

Figure 2 contains a subset of the paths returned from an execution of the algorithm for the word pair $< battle, war >$ merged into a graph. Intermediate nodes (words) and edges (higher level relations) provide a meaningful connection between the source and target nodes. Each path has an associated score which is the average of the semantic relatedness measures, which can serve as a ranking function to prioritize paths which are potentially more meaningful for a reasoning context. The output paths can be interpreted as an *abductive* process between the two words, providing a semantic justification under the structure of the relational graph. Table 2 shows examples of paths for lengths 2, 3 and 4. Nodes are connected through relations which were ommited.

The selectivity provided by the use of the distributional semantic relatedness measure as a node selection mechanism can be visualized in Figure 3 (A), where the distribution of the # of occurrences of the semantic relatedness values (y-axis) are shown in a logarithmic scale. The semantic relatedness values were collected during the navigation process for all comparisons performed during the execution of the experiment.

Table 2: Examples of semantically related paths returned by the algorithm.

Paths - Length 2	Paths - Length 3	Paths - Length 4
daughter, parent, **child**	**club**, team, play, **football**	**music**, song, single, record, **album**
episode, show, **series**	**chancellor**, politician, parliament, **government**	**soccer**, football, ball, major_league, **league**
country, continent, **europe**	**spouse**, family, wed, **married**	**author**, write, story, fiction, **book**
mayor, politician, **leader**	**actress**, act_in_play , go_on_stage, **actor**	**artist**, create_art, work_of_art, art, **paint**
video_game, computer_game, **software**	**film**, cinema, watch_movie, **movie**	**place**, locality, localize, locate, **location**
long, measure, **length**	**spouse**, wife, marriage, **husband**	**jew**, religion, ethnic_group, ethnic, **ethnicity**
husband, married_man, **spouse**	**aircraft**, fly, airplane, **pilot**	**war**, gun, rifle, firearm, **weapon**
artist, draw, **paint**	**country**, capital, national_city, **city**	**pilot**, fly, airplane, plane, **aircraft**
city, capital, **country**	**chancellor**, head_of_state, prime_minister, **government**	**chancellor**, member, cabinet, prime_minister, **government**
jew, temple, **religion**		

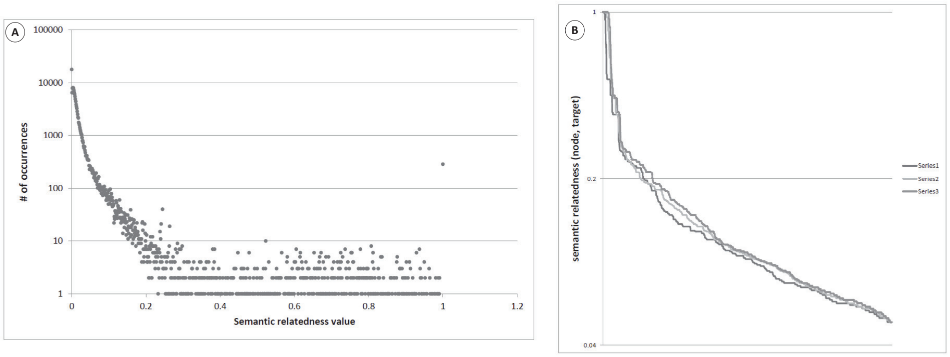

Fig. 3: # of occurrences for pairwise semantic relatedness values, computed by the navigational algorithm for the test collection (paths of length 2, 3, 4). Semantic relatedness values for nodes from distances 1, 2, 3 from the source: increasing semantic relatedness to the target.

The graph shows the discriminative efficiency of semantic relatedness, where just a tiny fraction of the entities in paths of length 2, 3, 4 are selected as semantically related to the target.

In Figure 3(B) the average increase on the semantic relatedness value as the navigation algorithm approaches the target is another pattern which can be observed. This smooth increase can be interpreted as an indicator of a meaningful path, where semantic relatedness value can serve as a heuristic to indicate a meaningful approximation from the target word. This is aligned with the increased selectivity of the Δ (semantic relatedness differential) criteria.

In the DNA algorithm, the semantic relatedness was used as a heuristic in a greedy search. The worst-case time complexity of a DFS is $O(b^l)$, where b is the branching factor and l is the depth limit. In this kind of search, the amount of performance improvement depends on the quality of the heuristic. In Table 1 we showed that as the depth limit increases, the selectivity of DNA ensures that the number of paths does not increase in the same amount. This indicates that the distributional semantic relatedness can be an effective heuristic when applied to the selection meaningful paths to be used in a reasoning process.

8 Related Work

Speer et al. (2008) introduced AnalogySpace, a hybrid distributional-relational model over ConceptNet using Latent Semantic Indexing. Cohen et al.(2009) proposes PSI, a distributional model that encodes predications produced by the SemRep system. The τ-Space distributional-relational model is similar to AnalogySpace and PSI. Differences in relation to these works are: (i) the supporting distributional model (τ-Space is based on Explicit Semantic Analysis), (ii) the use of the reference corpus (the τ-Space distributional model uses an independent large scale text corpora to build the distributional space, while PSI builds the distributional model based on the indexed triples), (iii) the application scenario (the τ-Space is evaluated under an open domain scenario while PSI is evaluated on the biomedical domain), (iv) the focus on evaluating the selectivity and ability to cope with incompleteness. Cohen et al.(2012) extends the discussion on the PSI to search over triple predicate pathways in a database of predications extracted from the biomedical literature by the SemRep system. Taking the data as a reference corpus, Novacek et al.(2011) build a distributional model which uses a PMI-based measure over the triple corpora. The approach was evaluated using biomedical semantic web data.

Freitas et al.(2011) introduces the τ-Space under the context of schema-agnostic queries over semantic web data. This work expands the discussion on the existing abstraction of the τ-Space, defined in [1], introducing the notion of selective reasoning process over a τ-Space.

Other works have concentrated on the relaxation of constraints for querying large KBs. SPARQLer (Kochut et al. [10]) is a SPARQL extension which allows query and retrieval of semantic associations (complex relationships) in RDF. The SPARQLer approach is based on the concept of path queries where users can specify graph path patterns, using regular expressions for example. The pattern matching process has been implemented as a hybrid of a bidirectional breadth-first search (BFS) and a simulation of a deterministic finite state automaton (DFA) created for a given path expression. Kiefer et al.(2007) introduce iSPARQL, a similarity join extension to SPARQL, which uses user-specified similarity functions (Levehnstein, Jaccard and TF/IDF) for potential assignments during query answering. Kiefer et al.(2007) considers that the choice of a best performing similarity measure is context and data dependent. Comparatively the approach described on this work focuses a semantic matching using distributional knowledge embedded in large scale corpora while iSPARQL focuses on the application of string similarity and SPARQLer on the manual specification of path patterns.

9 Conclusion

This work introduced a selective reasoning mechanism based on a distributional-relational semantic model which can be applied to heterogeneous commonsense KBs. The approach focuses on addressing the following problems: (i) providing a semantic selection mechanism for facts which are relevant and meaningful in a specific querying and reasoning context and (ii) allowing coping with information incompleteness in large KBs. The approach was evaluated using ConceptNet as a commonsense KB and ESA as the distributional model and achieved *high selectivity*, *high selectivity scalability* and *high* *accuracy in the selection of meaningful paths*. Distributional semantics was used as a principled mechanism to cope with information incompleteness. An estimation of information incompleteness for a real commonsense KB was provided and the suitability of distributional semantics to cope with it was verified. Future work will concentrate on improving the accuracy of the proposed approach by refining the distributional semantic model for the selective reasoning problem.

Acknowledgments. This research was supported in part by funding from the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight) and by the Irish Research Council. Joao C. P. da Silva is a CNPq Fellow - Science without Borders (Brazil).

References

1. Freitas, A., Curry, E., Oliveira, J. G., O'Riain, S., Distributional Structured Semantic Space for Querying RDF Graph Data. *International Journal of Semantic Computing*, 5(4), 433–462. (2011).
2. Freitas, A., Curry, E., O'Riain, S., A Distributional Approach for Terminology-Level Semantic Search on the Linked Data Web. *In Proc. 27th ACM Symp. On Applied Computing (SAC 2012)*, ACM Press. (2012).
3. Gabrilovich, E., Markovitch S., Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *In Proc. of the 20th Intl. Joint Conf. on Artificial Intelligence*, 1606–1611. (2007).
4. Kiefer, C., Bernstein, A., Stocker, M., The fundamentals of iSPARQL: A virtual triple approach for similarity-based semantic web tasks. *Lecture Notes in Computer Science*, vol. 4825, 295–295. (2007).
5. Turney, P.D., Pantel P., From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1), 141–188. (2010).
6. 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).
7. Cohen, T., Widdows, D., Schvaneveldt, R.W., Rindflesch, T.C.. Discovery at a Distance: Farther Journeys in Predication Space. *BIBM Workshops*, 218-225. (2012).
8. Cohen, T., Schvaneveldt, R.W., Rindflesch, T.C.. Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space. *T. AMIA Annu Symp Proc.*, 114-118. (2009).
9. 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).
10. Kochut, K., Janik, M., SPARQLeR: Extended SPARQL for semantic association discovery. *Lecture Notes in Computer Science*, 145-145. (2007).
11. Liu, H., Singh, P., ConceptNet A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal 22, 4*, 211-226. (2004).
12. Harris, Z., Distributional structure. *Word 10 (23)*, 146162. (1954).
13. 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).