Data & Knowledge Engineering 100 (2015) 211–225

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Approximate and selective reasoning on knowledge graphs: A distributional semantics approach

André Freitas [a,*], João C.P. da Silva [b,c], Edward Curry [b], Paul Buitelaar [b,d]

[a] Department of Computer Science and Mathematics, University of Passau, Dr-Hans-Kapfinger Strasse, 12, R.109, Passau, 94032, Germany
[b] Insight Centre for Data Analytics, National University of Ireland, Insight, IDA Buisiness Park, Lower Dangan, Galway, Ireland
[c] Computer Science Department, Federal University of Rio de Janeiro, Caixa-Postal 68530, CEP: 21941-916, Rio de Janeiro, RJ, Brazil
[d] College of Graduate Studies, University of South Africa, Pretoria, South Africa

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 24 June 2015
Accepted 25 June 2015
Available online 26 July 2015

Keywords:
Commonsense reasoning
Selective reasoning
Distributional semantics
Hybrid distributional-relation models
Semantic approximation

ABSTRACT

Tasks such as question answering and semantic search are dependent on the ability of querying and 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 and 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 and 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 selectivity 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 crowd-sourcing 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 over the KB. A complete KB would have the rule *'Every engineer has a degree'*, which would

* Corresponding author.
E-mail addresses: andre.freitas@uni-passau.de (A. Freitas), jcps.ufrj.br@gmail.com (J.C.P. da Silva).

http://dx.doi.org/10.1016/j.datak.2015.06.010
0169-023X/

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 (Fig. 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 Fig. 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 Fig. 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 Fig. 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 layers to the relational/logical 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 introduces natural language commonsense knowledge bases and briefly describes ConceptNet [11]; Section 3 provides an introduction on distributional semantics; Section 4 describes the τ-Space distributional-relational semantic model which is used for the selection reasoning mechanism; Section 5 defines the distributional representation for the commonsense *KB*; Section 6 describes the selective reasoning mechanism (*distributional navigational algorithm*); Section 7 provides an evaluation of the approach using explicit semantic analysis (ESA) as a distributional semantic model and ConceptNet [11] as KB; Section 8 describes related work and finally, Section 9 provides conclusions and future work.

2. Natural language commonsense knowledge bases (NLCS-KB)

More recent commonsense KBs such as ConceptNet are shifting from a logic representation framework to a natural language-based commonsense knowledge representation [11,14]. The motivation behind this change in perspective is to improve the scale

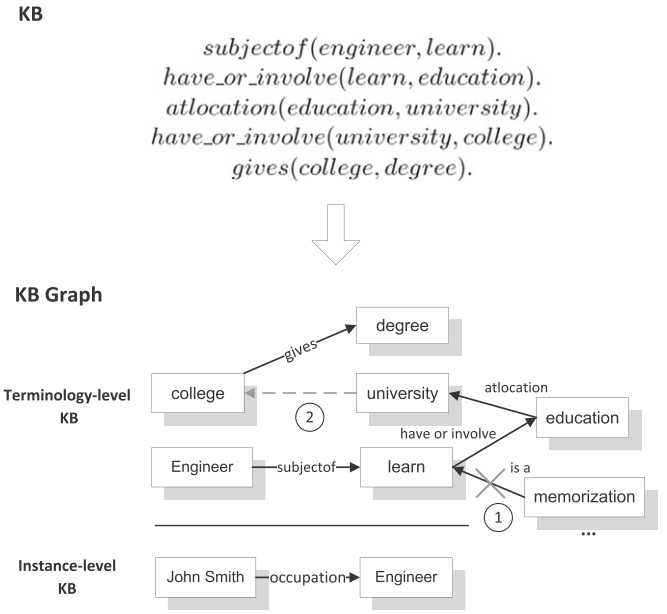

Fig. 1. (1) Selection of meaningful paths. (2) Coping with information incompleteness.

of acquiring and accessing commonsense knowledge. Natural language terms support the inheritance of the meaning of its cultural use, which can be contrasted to logical symbols, which have no a priori meaning outside its local context of definition [14].

Logical frameworks have a *stable and systematic way of evaluating and maintaining the truth of expressions* [14], where ambiguity and inconsistencies are removed during the construction of the KB. While logical KBs offer a precise framework for representing and inferring over knowledge, logical approaches present major scalability problems for the construction of large-scale commonsense KBs.

This work concentrates on the extension of natural language-based commonsense KBs with a distributional semantics framework, which aims at addressing incompleteness in the KB and provides a selective reasoning framework based on knowledge embedded in unstructured corpora. Despite the fact that this work concentrates on natural language commonsense KBs, the approach can be transported to logical KBs [15].

The ConceptNet natural language based KBs [14] is used to materialize the discussion. However, the proposed reasoning model can be transported to other NLCS-KBs with a similar structure.

ConceptNet is a large commonsense semantic network which is built from curated data and from data extracted from semi-structured and unstructured sources. ConceptNet can be seen as a labeled graph where the nodes represent natural language fragments which fall into three main classes: noun phrases, attributes, and activity phrases [14], and the labeled edges are described by an ontology of upper-level relations (Table 3). ConceptNet is stored as a set of triples of the form *(word1, relation, word2)*. For example, the sentences *"A telephone is used for communication"*, *"Linux is an operating system"* and *"You propose to a woman when you love her"* are respectively represented as the triples *(telephone,usedfor,communication)*, *(linux,isa,operating_system)* and *(propose_to_woman, motivatedbygoal,you_love_her)*. Note that in the last two cases we use short phrases as nodes instead words.

The English subset of ConceptNet 5 [11] is built from the following knowledge sources: (i) the Open Mind Commonse Sense website[1]; which collects manually curated commonsense knowledge; (ii) games with a purpose (e.g. verbosity); (iii) WordNet 3.0; (iv) information extraction on sources such as Wikitionary and Wikipedia; (v) structured information on Wikipedia.

The next sections describe the complementary distributional layer which is used to extend the semantics of the commonsense KB.

3. 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*.

DSMs are represented as a *vector space model*, where each dimension represents a *context pattern C* for the linguistic or data context in which the *target term 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 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.

Definition. Distributional semantic model (DSM): A distributional semantic model (DSM) is a tuple $(\mathcal{T}, \mathcal{C}, , \mathcal{W}, \mathcal{M}, d, \int)$, where:

- $\mathcal{T}$ is the set of *target words*, i.e. the words for which the DSM provides a contextual representation.
- $\mathcal{C}$ is the set of *context patterns* in which *T* co-occur.
- $\mathcal{R}$ is a *relation* between *T* and the set of context patterns *C*.
- $\mathcal{W}$ is the *context weighting scheme*.
- $\mathcal{M}$ is the *distributional matrix*, $\mathcal{T} \times \mathcal{C}$.
- *d* is the *dimensional reduction function*, $d : \mathcal{M} \rightarrow \mathcal{M}'$.
- *s* is the *distance measure*, between the vectors in $\mathcal{M}'$.

[1] http://openmind.media.mit.edu.

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

$$\vec{x} = \sum_{j=1}^{t} w_j \vec{c}_j \quad (1)$$

such that w is a context weighting scheme, i.e. a measure which define weights for the vector components.

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

In this work the distributional semantic model (DSM) used in the evaluation is *explicit semantic analysis* (ESA), which is briefly explained in the following section.

3.1. Explicit semantic analysis (ESA)

Explicit semantic analysis (ESA) [3] is a distributional semantic model based on Wikipedia. ESA represents the meaning of a text in a high-dimensional space of concepts derived from the Wikipedia text collection. In ESA, the distributional context window is defined by the Wikipedia article, where the context identifier is a Wikipedia article title/identifier.

A *universal ESA space* is created by building a vector space containing Wikipedia articles' document representations using the TF/IDF weighting scheme. In this space, each article is represented as a vector where each component is a weighted term present in the article. Once the space is built, a keyword query over the ESA space returns a list of ranked articles titles, which define a context vector associated with the query terms (where each vector component receives a relevance weight).

In the ESA model, the context is defined at the document level which defines a semantic model which captures both *syntagmatic* and *paradigmatic* relations, appropriate for the computation of a semantic relatedness measures for the schema-agnostic scenario. The coherence of the Wikipedia content discourse in the context of a Wikipedia article also influences the quality of the semantic relatedness measure.

The approach proposed by Gabrilovich and Markovitch also supports a simple compositionality model allows the composition of vectors for multi-word expressions, where the final concept is the centroid of the vectors representing the set of individual terms. The ESA semantic relatedness measure between two terms is calculated by computing the cosine similarity between two distributional vectors.

The link structure of the articles can be used for providing alternative or related expressions for the contexts (based on the extraction of anchor texts) and for the enrichment of the semantic model. The link structure can also work as a basis for dimensional reduction. Gabrilovich and Markovitch describe two levels of semantic interpretation models. *First-order* interpretation models are purely based on information present in the textual description of articles, while *second-order* models also include knowledge present in inter-article links.

Gabrilovich and Markovitch incorporate concept relations by boosting the weights of concepts linked from the top-k weight concepts. The authors apply a further generality filter, where only more general concepts extracted from links are considered. Generality is determined by the difference in the number of inlinks among two linked concepts. Since some articles are overly specific or are not completely developed, Gabrilovich and Markovitch prune some concepts based on heuristics of quality and relevance.

ESA has the following configuration parameters:

- $\mathcal{C}$ = Wikipedia article.
- $\mathcal{W}$ = TF/IDF.
- d = link-based pruning (optional).
- $\mathcal{S}$ = cosine.

4. 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. The *τ-Space* is built by extending a labeled graph G_{KB} with an associated distributional representation for each term used to name the graph labeled elements E. The hybrid distributional-structured representation enriches the semantic knowledge explicitly stated in the structured data with unstructured knowledge in the reference corpora.

The *τ-Space* is a *distributional-relational model (DRM)* which is defined as a tuple $(\mathcal{DSM}, \mathcal{DB}, \mathcal{RC}, \mathcal{F}, \mathcal{H})$, where:

Definition. Distributional-relational model (DRM):

- $\mathcal{DSM}$ is the *associated distributional semantic model.*
- $\mathcal{DB}$ is the *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 $\vec{\mathbf{e}}_\mathbf{i}$ in 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.

5. Embedding the commonsense KB into the τ-Space

We consider that a natural language commonsense knowledge base *KB* is formed by a set of *terms* $\{v_1, \cdots, v_n\}$ and a set of *relations* $\{r_1, \cdots, r_n\}$ between these terms, both represented as words or short phrases in natural language. Formally:

Definition. A commonsense knowledge base *KB* is defined by a *labeled digraph* $G_{KB}^{label} = (V, R, E)$, where $V = \{v_1, \cdots, v_n\}$ is a set of nodes, $R = \{r_1, \cdots, r_n\}$ 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:

Definition. Let *KB* be commonsense knowledge base and $G_{KB}^{label} = (V, R, E)$ be its labeled digraph representation. A simplified representation of *KB* is defined by a *digraph* $G_{KB} = (V', E')$, where $V' = V$ and $E' = \{(v_i, v_j) : (v_i, r, v_j) \in E\}$.

Each labeled element in the KB have a set of senses associated with the natural language expression used to describe the element. Adopting the simplified *KB* representation, the set of senses associated to a node can be defined as:

Definition (Node Senses). Let *KB* be a commonsense knowledge base represented by $G_{KB}^{label} = (V, R, E)$ and let *Sense* be the set of senses semantically supported by the *KB*. The function *NS*

$$NS : V \rightarrow \left(2^{Sense}\ \{\varnothing\}\right)$$

defines the node sense of all $v \in V$.

In the natural language KB, the set of senses are not explicitly represented. Each node in the graph have a set of senses which are semantically supported by the relationships with the neighboring nodes.

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}, as follows:

Definition. The vector representation of $G_{KB}^{label} = (V, R, E)$ in VS^{dist} is $\overrightarrow{\mathbf{G}}_{KB_{dist}}^{label} = (\overrightarrow{\mathbf{V}}_{dist}, \overrightarrow{\mathbf{R}}_{dist}, \overrightarrow{\mathbf{E}}_{dist})$ such that:

$$\overrightarrow{\mathbf{V}}_{dist} = \left\{ \overrightarrow{\mathrm{v}} : \overrightarrow{\mathrm{v}} = \sum_{i=1}^{t} u_i^v \overrightarrow{c}_i, \text{ for each } v \in V \right\} \tag{2}$$

$$\overrightarrow{\mathbf{R}}_{dist} = \left\{ \overrightarrow{\mathrm{r}} : \overrightarrow{\mathrm{r}} = \sum_{i=1}^{t} u_i^v \overrightarrow{c}_i, \text{ for each } r \in R \right\} \tag{3}$$

$$\overrightarrow{\mathbf{E}}_{dist} = \left\{ \overrightarrow{\mathrm{r}} - \overrightarrow{\mathrm{v}}, \overrightarrow{\mathrm{v}}_{\mathrm{j}} - \overrightarrow{\mathrm{r}} : \text{for each } \left(v_i, r, v_j\right) \in E \right\} \tag{4}$$

u_i^v and u_i^r are defined by the weighting scheme over the distributional model.[2]

After the *KB* is embedded into the distributional space, each node is enriched with a distributional representation. The distributional representation contains the set of contexts in which a word/term appears in the reference corpus, in all possible senses for that word/term according to the reference corpus (Fig. 2).

6. 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:

Definition. A semantic relatedness function $sr : VS^{dist} \times VS^{dist} \rightarrow [0, 1]$ is defined as:

$$sr\left(\overrightarrow{\mathbf{p}_1}, \overrightarrow{\mathbf{p}_2}\right) = \cos(\theta) = \overrightarrow{\mathbf{p}_1} \cdot \overrightarrow{\mathbf{p}_2}$$

A threshold $\eta \in [0, 1]$ can be used to establish the desired semantic relatedness between two vectors: $sr(\overrightarrow{\mathbf{p}_1}, \overrightarrow{\mathbf{p}_2}) > \eta$.

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

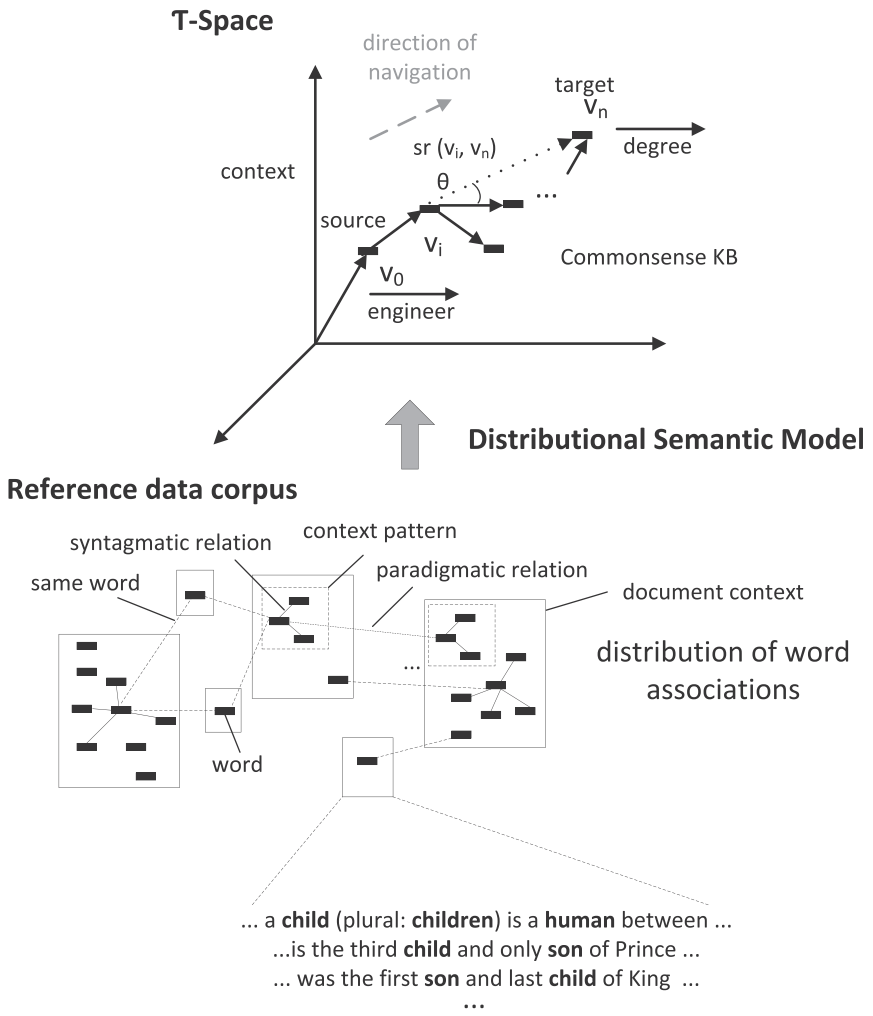

Fig. 2. Schematic depiction of the τ - Space embedding the commonsense KB graph into a distributional space.

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]. Multi-word phrases are handled by calculating the *centroid* between the context 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 (*lines 7–8*).

The stop condition is $sr(target, target) = 1$ (*lines 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:

$$rank(p) = \sum_{i=0}^{l} sr\left(\overrightarrow{t_i}, \overrightarrow{\mathbf{target}}\right) \tag{5}$$

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 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., $\Delta_{target}(t_k, n) > 0$.

Algorithm 1. Distributional navigation algorithm

INPUT

- *threshold*: η
- *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** $RankedPaths$

The navigation process of the distributional navigation algorithm can be interpreted as a node disambiguation process where nodes with senses which are strongly related to the *target* element are selected using the distributional relatedness measure. The distributional relatedness measure works as a word sense disambiguation mechanism by selecting nodes which have senses which have strong semantic relationships (expressed in the reference corpus) with the target node.

The *source* and *current* nodes also work as a contextual constraint (and as a disambiguation mechanism), which affects the selection of the nodes with compatible senses. As the *KB* does not explicitly represent node senses, this process works implicitly by selecting node terms which are semantically compatible with both source and target.

Definition (Sense Disambiguation). Let *KB* be a commonsense knowledge base represented by $G_{KB}^{label} = (V, R, E)$ and let η be a semantic relatedness threshold. Given two nodes $v_i, v_j \in V$, the sense disambiguation of v_i wrt v_j and η, is defined by the function

$$Disambiguation_{Sense}\left(v_i, v_j, \eta\right) = NS_{v_i, v_j} \subseteq NS(v_i) \cup UNS\left(v_j\right)$$

whenever $sr(v_i, v_j) > \eta$.

The mechanism defined above aims at maximizing the selection of a navigation path which is meaningful under the contextual constraints of the queries. The *disambiguation* function is implicitly defined by the composing the graph constraints between nodes and the navigation based on the distributional semantic relatedness computed against the neighboring nodes and the target words.

Additionally, the distributional semantic model supports detecting semantic relations between nodes which do not have an explicit relationship stated in the *KB*. This allows an extension of the distributional navigation algorithm to cope with *KB* incompleteness. The set of distributional unlabeled relations can be interpreted as an extension of the *KB* as defined below:

Definition (Distributional Relation). Let *KB* be a commonsense knowledge base represented by $G_{KB}^{label} = (V, R, E)$. The distributional relation r_η wrt a threshold η is defined as $\{(v_i, v_j) | v_i, v_j \in V; \forall\, r \in R, (v_i, r, v_j) \notin E$ and $sr(v_i, v_j) > \eta\}$.

Definition (Distributional Closure). Let *KB* be a commonsense knowledge base represented by $G_{KB}^{label} = (V, R, E)$. The distributional closure of *KB* wrt a threshold η is defined by the *labeled digraph* $G_{KB}^{label}(\eta) = (V, R \cup \{r_\eta\}, E \cup \{(v_i, r_\eta, v_j) | (v_i, v_j) \in r_\eta\})$.

7. Evaluation

7.1. 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. 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

Table 1
Number of triples per relation.

Number of triples	Number of relations
$=1$	45.311
$1 < x < 10$	11.804
$10 \leq x < 20$	906
$20 \leq x < 500$	790
≥ 500	50

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 2,252,338. that use 58,861 different relations. Most of the relations (45,311) have only one triple and only 50 relations appear in more than 500 triples. The results are summarized in Table 1. The distribution of the number of clauses per relation type is presented in Table 2.

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).[3] 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). Fifty-one 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.[4]

7.2. 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 3 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 lengths 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 3 shows the average selectivity for DNA. The variation of DNA with the Δ criteria, compared to DNA, provides a further selectivity improvement ($\phi =$ (# of spurious paths returned by DNA/# of spurious paths returned by DNA + Δ)) $\phi(length2) = 1$, $\phi(length3) = 0.49$, $\phi(length4) = 0.20$.

The results for each source, target pair can be found in Tables 4 and 5.

7.3. 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 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 two human annotators evaluated the relevance of the same set of paths. The paths which were in agreement between the two annotators were used in the experiment.

The accuracy of the algorithm for different path lengths can be found in Table 3. 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).

[3] http://www.sc.cit-ec.uni-bielefeld.de/qald-1.
[4] http://bit.ly/1p3PmHr.

Table 2
Top frequent relations in the ConceptNet.

Relation	Number of triples
instanceof	918.123
isa	201.710
hasproperty	120.961
subjectof	96.566
definedas	94.775
relatedto	88.922
directobjectof	87.946
usedfor	62.242
have_or_involve	49.967
atlocation	49.216
derivedfrom	40.403
capableof	38.811
synonym	34.974
hassubevent	27.366
hasprerequisite	25.160
causes	18.688
motivatedbygoal	16.178
be_in	15.143
be_near	12.744
be_not	11.777
receivesaction	11.095
hasa	10.048
partof	7.104

7.4. Addressing information incompleteness

This experiment measures the suitability of the distributional semantic relatedness measure to cope with KB incompleteness (gaps in the KB). Thirty-nine ⟨*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 neighboring entities. The distributional semantic search mechanism is equivalent to the computation of the semantic relatedness between the query and all entities (nodes) in the KB ($sr(\langle p_1, source\ target\rangle \langle p_n, source\ target\rangle)$). 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/# 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. number 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* (Table 6).

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 is expressed in the *incompleteness coefficient* which was found to be high, confirming the relevance of this problem under the context of reasoning over commonsense KBs.

Table 3
Selectivity and accuracy.

Path length	Average selectivity agorithm 1	% pairs of words resolved	Path accuracy
2	0.602	0.618	0.958
3	0.153	0.726	0.828
4	0.019	0.794	0.736

Table 4
The number of paths, selected paths and selectivity of the pair of terms (source,target) for paths with lengths of 2, 3 and 4.

	Length 2			Length 3			Length 4		
(target, source)	No. of path	No. of selected paths	Selectivity	No. of path	No. of selected paths	Selectivity	No. of path	No. of selected paths	Selectivity
chancellor–government	4	3	0.75	60	22	0.367	2060	184	0.089
battle–war	10	6	0.6	167	20	0.120	4472	76	0.017
daughter–child	15	4	0.267	327	11	0.034	9719	9	0.001
death–die	9	1	0.111	–	–	–	–	–	–
actress–actor	3	2	0.667	18	3	0.167	704	15	0.021
episode–series	1	1	1.000	–	–	–	282	2	0.007
single–song	1	1	1.000	14	4	0.286	354	22	0.062
country–europe	16	9	0.563	135	11	0.081	4823	27	0.006
mayor–leader	2	2	1.0	50	10	0.200	1401	41	0.029
high–elevation	1	1	1.000	16	2	0.125	263	3	0.011
video game–software	1	1	1.000	30	2	0.067	1090	8	0.007
music–album	14	4	0.286	411	28	0.068	18996	143	0.008
wife–spouse	3	2	0.667	32	5	0.156	1119	12	0.011
long–length	8	3	0.375	51	11	0.216	1192	37	0.031
movie–film	5	1	0.200	94	3	0.032	3524	1	0.000
husband–spouse	4	2	0.500	27	2	0.074	717	6	0.008
people–population	6	2	0.333	251	4	0.016	11501	22	0.002
artist–paint	2	2	1.000	52	4	0.077	2342	20	0.009
company–organization	25	11	0.440	700	36	0.051	30223	146	0.005
place–location	13	5	0.385	238	19	0.080	9672	91	0.009
city–country	23	9	0.391	588	36	0.061	19935	130	0.007
occupation–job	2	1	0.500	29	4	0.138	653	21	0.032
jew–religion	5	2	0.400	109	14	0.128	3601	64	0.018
soccer–ball	3	2	0.667	57	11	0.193	2004	45	0.022
war–weapon	7	2	0.286	92	11	0.120	2663	41	0.015
car–automobile	19	11	0.579	239	26	0.109	7250	91	0.013
pilot–aircraft	2	2	1.000	17	9	0.529	501	39	0.078
game–competition	6	3	0.500	155	13	0.084	5056	56	0.011
success–money	4	1	0.250	156	6	0.038	5850	43	0.007
country–moon	4	2	0.500	144	12	0.083	5129	63	0.012
spouse–married	–	–	–	3	3	1.000	41	11	0.268
chancellor–head	–	–	–	80	4	0.050	3231	11	0.003
european–europe	–	–	–	6	1	0.167	211	4	0.019
soccer–league	–	–	–	4	1	0.250	215	3	0.014
ruler–leader	–	–	–	20	4	0.200	832	22	0.026
author–book	–	–	–	97	8	0.082	3936	57	0.014
artist–song	–	–	–	87	5	0.057	3305	67	0.020
monarchy–government	–	–	–	1	1	1.000	90	5	0.056
jew–ethnicity	–	–	–	6	3	0.500	198	6	0.030
football–club	–	–	–	111	1	0.009	2888	1	0.000
university–professor	–	–	–	30	1	0.033	1004	6	0.006
player–instrument	–	–	–	–	–	–	3518	2	0.001
design–develop	–	–	–	–	–	–	1506	2	0.001

8. Analysis of the algorithm behavior

Fig. 3 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. Tables 7, 8 and 9 shows examples of paths for lengths 2, 3 and 4. Nodes are connected through relations which were omitted.

8.1. Navigation example

Consider that we want to find the paths of length 3 between the source *battle* and the target *war*, with threshold $\eta = 0.5$. Initially, we have the path $\langle t_0 \rangle = \langle battle \rangle$ and $sr(\overrightarrow{\mathbf{source}}, \overrightarrow{\mathbf{target}}) = sr(\overrightarrow{\mathbf{battle}}, \overrightarrow{\mathbf{war}}) = 0.064$.

The set of neighbors of the node *battle* in ConcepNet has 230 distinct elements related to it, such as *army*, *conflict*, *unit*, *Iraq*, *videogame* and *result in loss of life*. Among the 230 elements, the DNA algorithm selects the ones such that the semantic relatedness

Table 5
The number of paths, selected paths and selectivity of the pair of terms (source, target) for paths with lengths of 2, 3 and 4.

Pair	Length 2			Length 3			Length 4		
	No. of path	No. of selected paths	Selectivity	No. of path	No. of selected paths	Selectivity	No. of path	No. of selected paths	Selectivity
sex–metal	–	–	–	112	1	0.009	6690	7	0.001
man–source	–	–	–	230	2	0.009	11563	6	0.001
author–write	1	1	1.0	102	11	0.108	4011	80	0.020
married-spouse	1	1	1.0	6	1	0.167	332	3	0.009
wife–spouse	3	2	0.667	32	5	0.156	1119	12	0.011
head–chancellor	–	–	–	11	1	0.091	351	3	0.009
government–chancellor	2	1	0.5	14	2	0.143	262	5	0.019
war–battle	6	2	0.333	50	6	0.120	1645	31	0.019
child–daughter	4	2	0.500	75	6	0.080	2754	18	0.007
die–death	13	5	0.385	327	15	0.046	11109	31	0.003
mission–astronaut	1	1	1.0	–	–	–	–	–	–
song–single	9	4	0.444	221	14	0.063	9365	71	0.008
europe–country	4	1	0.250	79	1	0.013	2765	9	0.003
high–highest	2	1	0.500	–	–	–	119	2	0.017
software–video game	1	1	1.000	–	–	–	184	2	0.011
league–soccer	–	–	–	4	1	0.250	239	6	0.025
album–music	4	3	0.750	81	24	0.296	2445	122	0.050
film–movie	5	1	0.200	176	4	0.023	6990	7	0.001
leader–ruler	–	–	–	–	–	–	819	1	0.001
spouse–husband	1	1	1.000	6	2	0.333	35	1	0.029
population–people	8	4	0.500	257	20	0.078	10708	104	0.010
book–author	4	2	0.500	108	8	0.074	4529	13	0.003
paint–artist	3	2	0.667	77	15	0.195	2898	56	0.019
song–artist	6	3	0.500	122	25	0.205	3291	109	0.033
organization–company	8	4	0.500	104	14	0.135	4322	47	0.011
instrument–player	–	–	–	–	–	–	893	4	0.004
country–city	35	4	0.114	1030	28	0.027	39485	105	0.003
government–monarchy	1	1	1.000	22	5	0.227	447	18	0.040
religion–jew	2	2	1.000	20	2	0.100	637	9	0.014
ball–soccer	3	2	0.667	84	17	0.202	2529	62	0.025
club–football	4	1	0.250	111	6	0.054	4475	62	0.014
weapon–war	–	–	–	177	1	0.006	4104	16	0.004
develop–design	2	1	0.500	28	1	0.036	772	2	0.003
automobile–car	12	4	0.333	128	12	0.094	4134	52	0.013
aircraft–pilot	–	–	–	5	1	0.200	96	3	0.031
professor–university	5	4	0.800	115	29	0.252	3947	233	0.059
competition–game	5	3	0.600	105	14	0.133	3890	73	0.019
money–success	–	–	–	–	–	–	4667	2	0.000
moon–country	–	–	–	268	1	0.004	12497	5	0.000
window–religion	1	1	1.000	82	3	0.037	4504	31	0.007
cosmonaut–astronaut	2	2	1.000	3	1	0.333	–	–	–
job–occupation	4	1	0.250	–	–	–	–	–	–
length–long	2	2	1.000	–	–	–	–	–	–

wrt the target is greater than 0.05 ($sr(\overrightarrow{\mathbf{t}_1}, \overrightarrow{\mathbf{war}}) > 0.05$). In the examples, we have:

$$sr\left(\overrightarrow{\mathbf{army}}, \overrightarrow{\mathbf{war}}\right) = 0.135,$$

$$sr\left(\overrightarrow{\mathbf{conflict}}, \overrightarrow{\mathbf{war}}\right) = 0.172,$$

$$sr\left(\overrightarrow{\mathbf{unit}}, \overrightarrow{\mathbf{war}}\right) = 0.082,$$

$$sr\left(\overrightarrow{\mathbf{iraq}}, \overrightarrow{\mathbf{war}}\right) = 0.061.$$

Table 6
Semantically related entities returned.

Query: mayor–city–length 2
council, municipality, downtown, ward, incumbent borough, reelected, metropolitan, city elect, candidate, politician, democratic

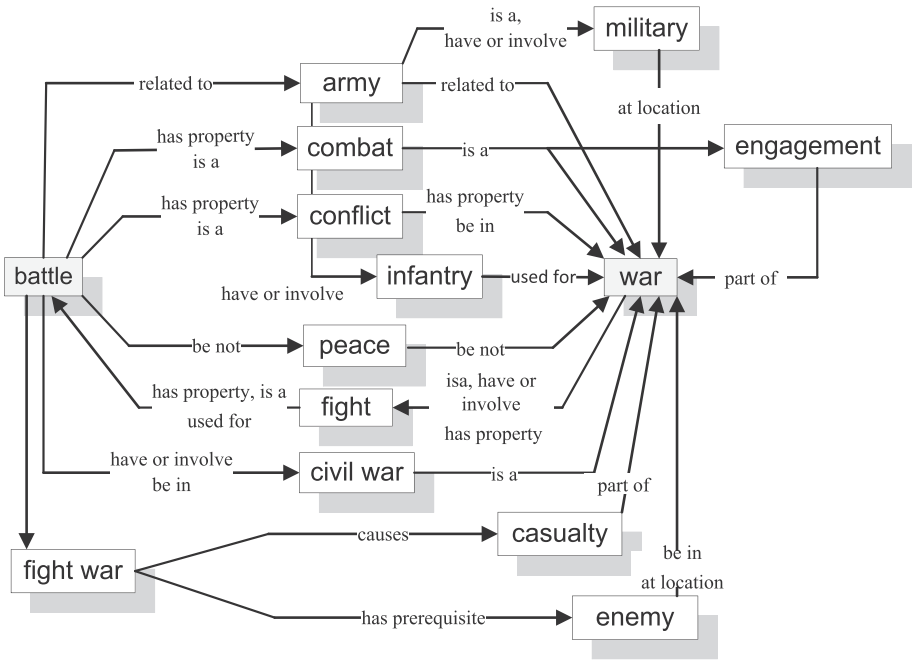

Fig. 3. Contextual (selected) paths between battle and war.

$$sr\left(\overrightarrow{\textbf{result in lossof life}}, \overrightarrow{\textbf{war}}\right) = 0.007.$$
$$sr\left(\overrightarrow{\textbf{videogame}}, \overrightarrow{\textbf{war}}\right) = 0.006.$$

and then ⟨*battle*, *army*⟩, ⟨*battle*, *conflict*⟩, ⟨*battle*, *unit*⟩ and ⟨*battle*, *iraq*⟩ are examples of the next paths to be explored. Note that we do not consider the relations involved since while some nodes have only one edge in ConceptNet, like (*battle, relatedto, army*), others have more than one, like (*battle, hasproperty, conflict*) and (*battle, isa, conflict*).

Continuing this process, from the 167 paths of length 3 in ConceptNet connecting *battle* and *war*, the DNA algorithm selected the following paths, with the corresponding score:

$score(\langle battle, fight\_war, military, war\rangle) = 1.679$
$score(\langle battle, fight\_war, army, war\rangle) = 1.658$
$score(\langle battle, fight\_war, soldier, war\rangle) = 1.628$
$score(\langle battle, fight\_war, peace, war\rangle) = 1.624$
$score(\langle battle, fight\_war, advance\_into\_battle, war\rangle) = 1.604$

Table 7
Examples of semantically related paths returned by the algorithm (paths–length 2).

Paths–length 2
daughter, parent, **child**
episode, show, **series**
country, continent, **europe**
mayor, politician, **leader**
video_game, computer_game, **software**
long, measure, **length**
husband, married_man, **spouse**
artist, draw, **paint**
city, capital, **country**
jew, temple, **religion**

Table 8
Examples of semantically related paths returned by the algorithm (paths–length 3).

Paths–length 3
club, team, play, **football**
chancellor, politician, parliament, **government**
spouse, family, wed, **married**
actress, act_in_play, go_on_stage, **actor**
film, cinema, watch_movie, **movie**
spouse, wife, marriage, **husband**
aircraft, fly, airplane, **pilot**
country, capital, national_city, **city**
chancellor, head_of_state, prime_minister, **government**

$score(\langle battle, fight\_war, attack, war\rangle) = 1.580$
$score(\langle battle, fight\_war, casualty, war\rangle) = 1.580$
$score(\langle battle, fight\_war, enemy, war\rangle) = 1.578$
$score(\langle battle, army, military, war\rangle) = 1.291$
$score(\langle battle, conflict, peace, war\rangle) = 1.273$
$score(\langle battle, peace, conflict, war\rangle) = 1.273$
$score(\langle battle, combat, conflict, war\rangle)$warS $= 1.248$
$score(\langle battle, army, soldier, war\rangle) = 1.240$
$score(\langle battle, army, military, war\rangle) = 1.238$
$score(\langle battle, army, infantry, war\rangle) = 1.190$
$score(\langle battle,peace, hostility, war\rangle) = 1.181$
$score(\langle battle, army, general, war\rangle) = 1.139$
$score(\langle battle, unit, infantry, war\rangle) = 1.137$
$score(\langle battle, combat, engagement, war\rangle) = 1.127$
$score(\langle battle, iraq, history, war\rangle) = 1.118$

From these paths, if we consider the paths that the semantic relatedness measure increases along the path (Δ_{target} heuristics), only four paths among these will be selected:

- $\langle battle, army, military, war\rangle$: since$[sr(\overrightarrow{\mathbf{army}}, \overrightarrow{\mathbf{war}}), sr(\overrightarrow{\mathbf{military}}, \overrightarrow{\mathbf{war}})] = [0.135, 0.156]$
- $\langle battle, peace, conflict, war\rangle$: since $[sr(\overrightarrow{\mathbf{peace}}, \overrightarrow{\mathbf{war}}), sr(\overrightarrow{\mathbf{conflict}}, \overrightarrow{\mathbf{war}})] = [0.101, 0.172]$
- $\langle battle, combat, conflict, war\rangle$): since $[sr(\overrightarrow{\mathbf{combat}}, \overrightarrow{\mathbf{war}}), sr(\overrightarrow{\mathbf{conflict}}, \overrightarrow{\mathbf{war}})] = [0.076, 0.172]$
- $\langle battle, unit, military, war\rangle$): since $[sr(\overrightarrow{\mathbf{unit}}, \overrightarrow{\mathbf{war}}), sr(\overrightarrow{\mathbf{military}}, \overrightarrow{\mathbf{war}})] = [0.082, 0.156]$

The selectivity provided by the use of the distributional semantic relatedness measure as a node selection mechanism can be visualized in Fig. 4, where the distribution of the number 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. The graph shows the discriminative efficiency of semantic relatedness, where just a tiny fraction of the entities in paths of lengths 2, 3, 4 are selected as semantically related to the target.

In Fig. 5 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

Table 9
Examples of semantically related paths returned by the algorithm (paths–length 4).

Paths–length 4
music, song, single, record, **album**
soccer, football, ball, major_league, **league**
author, write, story, fiction, **book**
artist, create_art, work_of_art, art, **paint**
place, locality, localize, locate, **location**
jew, religion, ethnic_group, ethnic, **ethnicity**
war, gun, rifle, firearm, **weapon**
pilot, fly, airplane, plane, **aircraft**
chancellor, member, cabinet, prime_minister, **government**

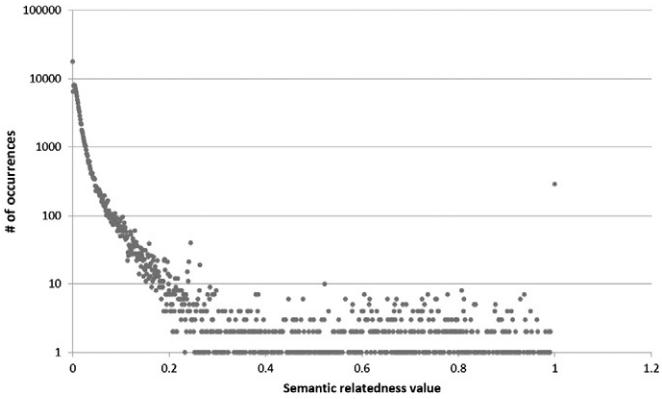

Fig. 4. Number of occurrences for pairwise semantic relatedness values, computed by the navigational algorithm for the test collection (paths of lengths 2, 3, 4).

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 3 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.

9. Related work

Speer et al. [6]) introduced AnalogySpace, a hybrid distributional-relational model over ConceptNet using Latent Semantic Indexing. Cohen et al. [8] 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. [7] 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

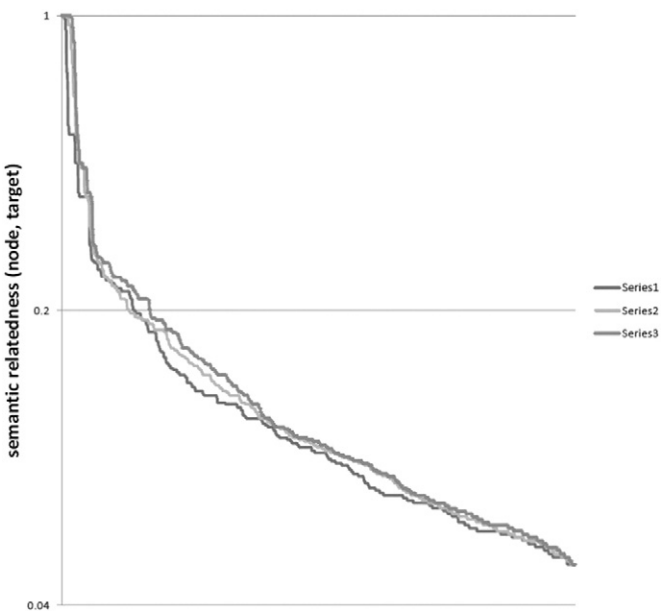

Fig. 5. Semantic relatedness values for nodes from distances 1, 2, 3 from the source: increasing semantic relatedness to the target.

system. Taking the data as a reference corpus, Novacek et al. [9] 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. [1] 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 [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. [4] 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. [4] 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.

10. 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.

References

[1] A. Freitas, E. Curry, J.G. Oliveira, S. O'Riain, Distributional structured semantic space for querying RDF graph data, Int. J. Semant. Comput. 5 (4) (2011) 433–462.
[2] A. Freitas, E. Curry, O'RiainS. , A distributional approach for terminology-level semantic search on the linked data web, 27th ACM Symp. on Applied Computing (SAC 2012). ACM Press, 2012.
[3] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, Proc. of the 20th Intl. Joint Conference on Artificial Intelligence 2007, pp. 1606–1611.
[4] C. Kiefer, A. Bernstein, M. Stocker, The fundamentals of iSPARQL: a virtual triple approach for similarity-based semantic web tasks, Lect. Notes Comput. Sci 4825 (2007) 295-295.
[6] R. Speer, C. Havasi, H. Lieberman, AnalogySpace: reducing the dimensionality of common sense knowledge, Proceedings of the 23rd International conference on, Artificial Intelligence 2008, pp. 548–553.
[7] T. Cohen, D. Widdows, R.W. Schvaneveldt, T.C. Rindflesch, Discovery at a distance: farther journeys in predication space, BIBM Workshops2012 218–225.
[8] T. Cohen, R.W. Schvaneveldt, T.C. Rindflesch, Predication-based semantic indexing: permutations as a means to encode predications in semantic space, *T. AMIA Annu Symp Proc*2009 114–118.
[9] V. Novacek, S. Handschuh, S. Decker, Getting the meaning right: a complementary distributional layer for the web semantics, Proceedings of ISWC 2011, pp. 504–519.
[10] K. Kochut, M. Janik, SPARQLeR: extended SPARQL for semantic association discovery, Lect. Notes Comput. Sci (2007) 145-145.
[11] H. Liu, P. Singh, ConceptNet: a practical commonsense reasoning tool-kit, BT Technol. J. 22 (4) (2004) 211–226.
[12] Z. Harris, Distributional structure, Word 10 (23) (1954) 146–162.
[14] H. Liu, P. Singh, Commonsense reasoning in and over natural language, Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES-2004), 2004.
[15] J. Pereira da Silva, A. Freitas, Towards an approximative ontology-agnostic approach for logic programs, Proceedings of the Eighth International Symposium on Foundations of Information and Knowledge Systems (FolKS), 2014.