Semantic Web 1 (2016) 1–5
IOS Press

Extracting common sense knowledge via triple ranking using supervised and unsupervised distributional models

Editor(s): Claudia d'Amato, University of Bari, Italy; Agnieszka Lawrynowicz, Poznan University of Technology, Poland; Jens Lehmann, University of Bonn and Fraunhofer IAIS, Germany
Solicited review(s): Dagmar Groman, TU Dresden, Germany; Ziqi Zhang, University of Sheffield, United Kingdom; Jedrzej Potoniec, Poznan University of Technology, Poland

Soufian Jebbara [a,*], Valerio Basile [b,**], Elena Cabrio [b], and Philipp Cimiano [a]
[a] *CITEC, Bielefeld University, Inspiration 1, 33619, Bielefeld, Germany*
E-mail: {sjebbara,cimiano}@cit-ec.uni-bielefeld.de
[b] *Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France*
E-mail: valerio.basile@inria.fr; elena.cabrio@unice.fr

Abstract.
In this paper we are concerned with developing information extraction models that support the extraction of common sense knowledge from a combination of unstructured and semi-structured datasets. Our motivation is to extract manipulation-relevant knowledge that can support robots' action planning. We frame the task as a relation extraction task and, as proof-of-concept, validate our method on the task of extracting two types of relations: locative and instrumental relations. The locative relation relates objects to the prototypical places where the given object is found or stored. The second instrumental relation relates objects to their prototypical purpose of use. While we extract these relations from text, our goal is not to extract specific textual mentions, but rather, given an object as input, extract a ranked list of locations and uses ranked by 'prototypicality'. We use distributional methods in embedding space, relying on the well-known skip-gram model to embed words into a low-dimensional distributional space, using cosine similarity to rank the various candidates. In addition, we also present experiments that rely on the vector space model NASARI, which compute embeddings for disambiguated concepts and are thus semantically aware. While this distributional approach has been published before, we extend our framework by additional methods relying on neural networks that learn a score to judge whether a given candidate pair actually expresses a desired relation. The network thus learns a scoring function using a supervised approach. While we use a ranking-based evaluation, the supervised model is trained using a binary classification task. The resulting score from the neural network and the cosine similarity in the case of the distributional approach are both used to compute a ranking.

We compare the different approaches and parameterizations thereof on the task of extracting the above mentioned relations. We show that the distributional similarity approach performs very well on the task. The best performing parameterization achieves an NDCG of 0.913, a Precision@1 of 0.400 and a Precision@3 of 0.423. The performance of the supervised learning approach, in spite of having being trained on positive and negative examples of the relation in question, is not as good as expected and achieves an NCDG of 0.908, a Precision@1 of 0.454 and a Precision@3 of 0.387, respectively.

Keywords: Relation Extraction, Distributional Semantics, Supervised Learning, Commonsense Knowledge

1. Introduction

Embodied intelligent systems such as robots require world knowledge to reason on top of their perception

*Corresponding author. E-mail: sjebbara@cit-ec.uni-bielefeld.de
**Corresponding author. E-mail: valerio.basile@inria.fr

of the world in order to decide which actions to take. Consider the example of a robot having the task to tidy up an apartment by storing all objects in their appropriate place. In order to perform this task, a robot would need to understand where the "correct" or at least the "prototypical" location for each object is in order to come up with an overall plan on which actions to perform to reach the goal of having each object stored in its corresponding location.

In general, in manipulating objects, robots might have questions such as the following:

- Where should a certain object typically be stored?
- What is this object typically used for?
- Do I need to manipulate a certain object with care?

The answers to these questions require common sense knowledge about objects, in particular prototypical knowledge about objects that, in absence of abnormal situations or specific contextual conditions or preferences, can be assumed to hold.

In this article, we are concerned with extracting such common sense knowledge from a combination of unstructured and semi-structured data. We are in particular interested in extracting default knowledge, that is prototypical knowledge comprising relations that typically hold in 'normal' conditions [42]. For example, given no other knowledge, in a normal situation, we could assume that milk is typically stored in the kitchen, or more specifically in the fridge. However, if a person is currently having breakfast and eating cornflakes at the table in the living room, then the milk might also be temporarily located in the living room. In this sense, inferences about the location of an object are to be regarded as non-monotonic inferences that can be retracted given some additional knowledge about the particular situation. We model such default, or prototypical, knowledge through a degree of prototypicality, that is, we do not claim that the kitchen is 'the prototypical location' for the milk, but instead we model that the degree of prototypicality for the kitchen being the default location for the milk is very high. This leads naturally to the attempt to computationally model this degree of prototypicality and rank locations or uses for each object according to this degree of prototypicality. We attempt to do so following two approaches. On the one hand, we follow a distributional approach and approximate the degree of prototypicality by the cosine similarity measure in a space into which entities and locations are embedded. We experiment with different distributional spaces and show that both semantic vector spaces as considered within the NASARI approach as well as embedded word representations computed on unstructured texts as produced by predictive language models such as skip-grams provide already a reasonable performance on the task. A linear combination of both approaches has the potential to improve upon both approaches in isolation. We have presented this approach before including empirical results for the `locatedAt` relation mentioned above in previous work [5]. As a second approach to approximate the degree of prototypicality, we use a machine learning approach trained on positive and negative examples using a binary classification scheme. The machine learning approach is trained to produce a score that measures the compatibility of a given pair of object and location/use in terms of their prototypicality. We compare these two approaches in this paper, showing that the machine learning approach does not perform as well as expected. Contrary to our intuitions, the unsupervised approach relying on cosine similarity in embedding space represents a very strong baseline difficult to beat.

The prototypical knowledge we use to train and evaluate the different methods is on the one hand based on a crowdsourcing experiment in which users had to explicitly rate the prototypicality of a certain location for a given object. On the other hand, we also use extracted relations from ConceptNet and the SUN database [70]. Objects as well as candidate locations, or candidate uses in the case of the instrumental relation, are taken from DBpedia. While we apply our models to known objects, locations and uses, our model could also be applied to candidate objects, locations and uses extracted from raw text.

We have different motivations for developing such an approach to extract common sense knowledge from unstructured and semi-structured data.

First, from the point of view of cognitive robotics [40] and cognitive development, acquiring common sense knowledge requires many reproducible and similar experiences from which a system can learn how to manipulate a certain object. Some knowledge can arguably even not be acquired by self experience as relevant knowledge also comprises the mental properties that humans ascribe to certain objects. Such mental properties that are not intrinsic to the physical appearance of the object include for instance the intended use of an object. There are thus limits to what can be learned from self-guided experience with an object. In fact, several scholars have emphasized the importance of cultural learning, that is of a more direct trans-

mission of knowledge via communication rather than self-experience. With our approach we are simulating such a cultural transmission of knowledge by allowing cognitive systems, or machines in our case, to acquire knowledge by 'reading' texts. Work along these lines has, for instance, tried to derive plans on how to prepare a certain dish by machine reading descriptions of household tasks written for humans that are available on the Web [67]. Other work has addressed the acquisition of scripts from the Web [55].

Second, while there has been a lot of work in the field of information extraction on extracting relations, the considered relations differ from the ones we investigate in this work. Standard relations considered in relation extraction are: *is-a, part-of, succession, reaction, production* [53,11] or *relation, parent/child, founders, directedBy, area_served, containedBy, architect, etc.* [57], or *albumBy, bornInYear, currencyOf, headquarteredIn, locatedIn, productOf, teamOf* [6]. The literature so far has focused on relations that are of a factual nature and explicitly mentioned in the text. In contrast, we are concerned with relations that are *i)* typically not mentioned explicitly in text, and *ii)* they are not of a factual nature, but rather represent default or prototypical knowledge. These are thus quite different tasks.

We present and compare different approaches to collect manipulation-relevant knowledge by leveraging textual corpora and semi-automatically extracted entity pairs. The extracted knowledge is of symbolic form and represented as a set of `(Subject, Relation, Object)` triples. While this knowledge is not physically grounded [24], this model can still help robots or other intelligent systems to decide on how to act, support planning and select the appropriate actions to manipulate a certain object.

The paper is structured as follows: In Section 2, we discuss related work from the fields of relation extraction, knowledge base population and knowledge bases for robotics. In Section 3, we describe our approach to relation extraction in general and continue by introducing two models based on semantic relatedness as a ranking measure. These two models have been described in earlier work [5] and are described here again for the sake of completeness and due to the fact that we compare this previous work to a novel approach we introduce in Section 3.3. The model introduced in Section 3.3 is a supervised model that is trained to extract arbitrary relations. Afterwards, in Section 4, we present our datasets that are used for training and evaluating the proposed models. We evaluate and compare all models in Section 5, showing that both unsupervised approaches and their combination perform very well on the task, outperforming two naive baselines. The supervised approach, while being superior with respect to Precision@1, does not show any clear benefit compared to the unsupervised approach, a surprising result.

In Section 6, we exploit insights gained from the evaluation to populate a knowledge base of manipulation-relevant data using the presented semi-automatic methods. Finally, in Section 7, we summarize our results and discuss directions for future work.

2. Related Work

Our work relates to the four research lines discussed below, namely: *i)* machine reading, *ii)* supervised relation extraction, *iii)* encoding common sense knowledge in domain-independent ontologies and knowledge bases, and *iv)* grounding of knowledge from the perspective of cognitive linguistics.

The machine reading paradigm. In the field of knowledge acquisition from the Web, there has been substantial work on extracting taxonomic (e.g. hypernym), part-of relations [23] and complete qualia structures describing an object [14]. Quite recently, there has been a focus on the development of systems that can extract knowledge from any text on any domain (the open information extraction paradigm [21]). The DARPA Machine Reading Program [1] aimed at endowing machines with capabilities for lifelong learning by automatically reading and understanding texts (e.g. [20]). While such approaches are able to quite robustly acquire knowledge from texts, these models are not sufficient to meet our objectives since: *i)* they lack visual and sensorimotor grounding, *ii)* they do not contain extensive object knowledge. While the knowledge extracted by our approach presented here is also not sensorimotorically grounded, we hope that it can support planning of tasks involving object manipulation. Thus, we need to develop additional approaches that can harvest the Web to learn about usages, appearance and functionality of common objects. While there has been some work on grounding symbolic knowledge in language [51], so far there has been no serious effort to compile a large and grounded object knowledge base that can support cognitive systems in understanding objects.

Supervised Relation Extraction. While machine reading attempts to acquire general knowledge by reading texts, other works attempt to extract specific relations using classifiers trained in a supervised approach using labeled data. A training corpus in which the relation of interest is annotated is typically assumed (e.g. [11]). Another possibility is to rely on the so called *distant supervision* approach and use an existing knowledge base to bootstrap the process by relying on triples or facts in the knowledge base to label examples in a corpus (e.g. [28,29,28,64]). Some researchers have modeled relation extraction as a matrix decomposition problem [57]. Other researchers have attempted to train relation extraction approaches in a bootstrapping fashion, relying on knowledge available on the Web, e.g. [7].

Recently, scholars have tried to build models that can learn to extract generic relations from the data, rather than a set of pre-defined relations (see [38] and [8]). Related to these models are techniques to predict triples in knowledge graphs by relying on the embedding of entities (as vectors) and relations (as matrices) in the same distributional space (e.g. TransE [10] and TransH [69]). Similar ideas were tested in computational linguistics in the past years, where relations and modifiers are represented as tensors in the distributional space [3,18].

Ontologies and KB of common sense knowledge. DBpedia[1] [36] is a large-scale knowledge base automatically extracted from the infoboxes of Wikipedia. Besides its sheer size, it is attractive for the purpose of collecting general knowledge given the one-to-one mapping with Wikipedia (allowing us to exploit the textual and structural information contained in there) and its position as the central hub of the Linked Open Data cloud.

YAGO [63] is an ontology automatically extracted from WordNet and Wikipedia. YAGO extracts facts from the category system and the infoboxes of Wikipedia, and combines these facts with taxonomic relations derived from WordNet. Despite its high coverage, for our goals, YAGO suffers from the same drawbacks as DBpedia, i.e., a lack of knowledge about common objects, that is, about their purpose, functionality, shape, prototypical location, etc.

ConceptNet[2] [39] is a semantic network containing lots of things computers should know about the world. However, we cannot integrate ConceptNet directly in our pipeline because of the low coverage of the mapping with DBpedia— of the 120 DBpedia entities in our gold standard (see Section 4) only 23 have a correspondent node in ConceptNet.

NELL (Never Ending Language Learning) is the product of a continuously-refined process of knowledge extraction from text [49]. Although NELL is a large-scale and quite fine-grained resource, there are some drawbacks that prevent it to be effectively used as a commonsense knowledge base. The inventory of predicates and relations is very sparse, and categories (including many objects) have no predicates.

OpenCyC[3] [37] attempts to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning.

Several projects worldwide have attempted to develop knowledge bases for robots through which knowledge, e.g. about how to manipulate certain objects, can be shared among many robots. Examples of such platforms are the RoboEarth project [68], RoboBrain [59] or KnowRob [66].

While the above resources are without doubt very useful, we are interested in developing an approach that can extract new knowledge leveraging text corpora, complementing the knowledge contained in ontologies and knowledge bases such as the ones described above.

Grounded Knowledge and Cognitive Linguistics Many scholars have argued that, from a cognitive perspective, knowledge needs to be grounded [24] as well as modality-specific to support simulation, a mental activity that is regarded as ubiquitous in cognitive intelligent systems [4]. Other seminal work has argued that cognition is categorical [25,26] and that perceptual and cognitive reasoning rely on schematic knowledge. In particular, there has been substantial work on describing the schemas by which we perceive and understand spatial knowledge [65].

The knowledge we have gathered is neither grounded nor schematic, nor modality-specific in the above senses, but rather amodal and symbolic. This type of knowledge is arguably useful in high-level planning but clearly is not sufficient to support simulation or event action execution. Developing models by which natural language can be grounded in action has been

[1] http://dbpedia.org
[2] http://conceptnet5.media.mit.edu/
[3] http://www.opencyc.org/ as RDF representations: http://sw.opencyc.org/

the concern of other authors, e.g. Misra et al. [47] as well as Bollini et al. [9]. Some work has considered extracting spatial relations in natural language input [33]. Differently from the above mentioned works, we are neither interested in interpreting natural language with respect to grounded action representations nor in extracting spatial relations from a given sentence. Rather, our goal is to extract prototypical common sense background knowledge from large corpora.

3. Extraction of Relations by a Ranking Approach based on Distributional Representations

This section presents our framework to extract relations between pairs of entities for the population of a knowledge base of manipulation-relevant data. We frame the task of relation extraction between entities as a ranking problem as it gives us great flexibility in generating a knowledge base that balances between coverage and confidence. Given a set of triples (s, r, o), where s is the subject entity, r the relation (or predicate) and o the object entity[4], we want to obtain a ranking of these triples. The produced ranking of triples should reflect the degree of prototypicality of the objects with respect to the respective subjects and relations.

Our general approach to produce these rankings is to design a scoring function $f(s, r, o)$ that assigns a score to each triple, depending on s, r, and o. The scoring function is designed in such a way that prototypical triples are assigned a higher score than less prototypical triples. Sorting all triples by their respective scores produces the desired ranking. With a properly chosen function $f(s, r, o)$, it is possible to extract relations between entities to populate a knowledge base. This is achieved by scoring candidate triples and inserting or rejecting them based on their respective scores, e.g. if the score is above a certain threshold.

In this work, we present different scoring functions and evaluate them in the context of building a knowledge base of common sense triples. All of our proposed approaches rely on distributional representations of entities (and words). We investigate different vector representations and scoring functions, all with different strengths and weaknesses. In the following, for the sake of making the article self-contained, we give a short introduction to distributional representations.

Word space models (or distributional space models, or word vector spaces) are abstract representations of the meaning of words, encoded as vectors in a high-dimensional space. Traditionally, a word vector space is constructed by counting *cooccurrences* of pairs of words in a text corpus, building a large square n-by-n matrix where n is the size of the vocabulary and the cell i, j contains the number of times the word i has been observed in cooccurrence with the word j. The i-th row in a cooccurrence matrix is an n-dimensional vector that acts as a *distributional* representation of the i-th word in the vocabulary. The similarity between two words is geometrically measurable with a metric such as the cosine similarity, defined as the cosine of the angle between two vectors:

$$similarity(\vec{x}, \vec{y})_{cos} = \frac{\vec{x} \cdot \vec{y}}{\| \vec{x} \| \| \vec{y} \|}$$

This is the key point to linking the vector representation to the idea of semantic relatedness, as the *distributional hypothesis* states that "words that occur in the same contexts tend to have similar meaning" [27]. Several techniques can be applied to reduce the dimensionality of the cooccurrence matrix. Latent Semantic Analysis [34], for instance, uses Singular Value Decomposition to prune the less informative elements while preserving most of the topology of the vector space, and reducing the number of dimensions to 100-500.

Recently, neural network based models have received increasing attention for their ability to compute dense, low-dimensional representations of words. To compute such representation, called *word embeddings*, several models rely on huge amounts of natural language texts from which a vector representation for each word is learned by a neural network. Their representations of the words are therefore based on *prediction* as opposed to *counting* [2].

Vector spaces created on word distributional representations have been successfully proven to encode word similarity and relatedness relations [54,56,15], and word embeddings have proven to be a useful feature in many natural language processing tasks [16,35, 19] in that they often encode semantically meaningful information of a word.

We argue that it is possible to extract interaction-relevant relations between entities, e.g. (`Object, locatedAt, Location`), using appropriate entity

[4] Here we use the terminology `subject` and `object` from the Semantic Web literature instead of the terminology `head` and `tail` that is typically found in relation extraction literature.

vectors and the cosine similarity since the domain and range of the considered relations are sufficiently narrow. In these cases, the semantic relatedness might be a good indicator for a relation.

3.1. Ranking by Cosine Similarity and Word Embeddings

In the beginning of this section, we motivated the use of distributional representations for the extraction of relations in order to populate a database of common sense knowledge. As outlined, we frame the relation extraction task as a ranking problem of triples (s, r, o) and score them based on a corresponding set of vector representations $\boldsymbol{V}$ for subject and object entities.

In this section, we propose a neural network-based word embedding model to obtain distributional representations of entities. By using the relation-agnostic cosine similarity[5] as our scoring function $f(s,r,o) = similarity_{cos}(\vec{v}_s, \vec{v}_o)$, with $\vec{v}_s, \vec{v}_o \in \boldsymbol{V}$, we can interpret the vector similarity as a measure of semantic relatedness and thus as an indicator for a relation between the two entities.

Many word embedding methods encode useful semantic and syntactic properties [32,48,44] that we leverage for the extraction of prototypical knowledge. In this work, we restrict our experiments to the skip-gram method [43]. The objective of the skip-gram method is to learn word representations that are useful for predicting context words. As a result, the learned embeddings often display a desirable linear structure [48,44]. In particular, word representations of the skip-gram model often produce meaningful results using simple vector addition [44]. For this work, we trained the skip-gram model on a corpus of roughly 83 million Amazon reviews [41].

Motivated by the compositionality of word vectors, we derive vector representations for the entities as follows: considering a DBpedia entity[6] such as `Public_toilet`, we obtain the corresponding label and clean it by removing parts in parenthesis, if any, convert it to lower case, and split it into its individual words. We retrieve the respective word vectors from our pre-trained word embeddings and sum them to obtain a single vector, namely, the vector representation of the entity: $\vec{v}_{Public\_toilet} = \vec{v}_{public} + \vec{v}_{toilet}$. The generation of entity vectors is trivial for "single-word" entities, such as `Cutlery` or `Kitchen`, that are already contained in our word vector vocabulary. In this case, the entity vector is simply the corresponding word vector. By following this procedure for every entity in our dataset, we obtain a set of entity vectors $\boldsymbol{V}_{sg}$, derived from the original skip-gram word embeddings. With this derived set of entity vector representations, we can compute a score between pairs of entities based on the chosen scoring function, the cosine vector similarity[7]. Using the example of `locatedAt`-pairs, this score is an indicator of how typical the location is for the object. Given an object, we can create a ranking of locations with the most prototypical location candidates at the top of the list (see Table 1). We refer to this model henceforth as `SkipGram/Cosine`.

Table 1
Locations for a sample object, extracted by computing cosine similarity on skip-gram-based vectors.

Object	Location	Cos. Similarity
Dishwasher	Kitchen	.636
	Laundry_room	.531
	Pantry	.525
	Wine_cellar	.519

3.2. Ranking by Cosine Similarity and Semantically-Aware Entity Representations

Vector representations of words (Section 3.1) are attractive since they only require a sufficiently large text corpus with no manual annotation. However, the drawback of focusing on words is that a series of linguistic phenomena may affect the vector representation. For instance, a polysemous word as *rock* (stone, musical genre, metaphorically strong person, etc.) is represented by a single vector where all the senses are conflated.

NASARI [12], a resource containing vector representations of most of DBpedia entities, solves this problem by building a vector space of concepts. The NASARI vectors are actually distributional representations of the entities in BabelNet [52], a large multilingual lexical resource linked to WordNet, DBpedia, Wiktionary and other resources. The NASARI approach collects cooccurrence information of concepts

[5] We also experimented with APSyn [58] as an alternative similarity measure which, unfortunately, did not work well in our scenario.

[6] For simplicity, we only use the local parts of the entity URI, neglecting the namespace `http://dbpedia.org/resource/`

[7] For any entity vector that can not be derived from the word embeddings due to missing vocabulary, we assume a similarity of -1 to every other entity.

from Wikipedia and then applies a cluster-based dimensionality reduction. The context of a concept is based on the set of Wikipedia pages where a mention of it is found. As shown by Camacho-Collados et al. [12], the vector representations of entities encode some form of semantic relatedness, with tests on a sense clustering task showing positive results. Table 2 shows a sample of pairs of NASARI vectors together with their pairwise cosine similarity ranging from -1 (opposite direction, i.e. unrelated) to 1 (same direction, i.e. related).

Table 2
Examples of cosine similarity computed on NASARI vectors.

	Cherry	Microsoft
Apple	.917	.325
Apple_Inc.	.475	.778

Following the hypothesis put forward in the beginning of this section, we focus on the extraction of interaction-relevant relations by computing the cosine similarities of entities. We exploit the alignment of BabelNet with DBpedia, thus generating a similarity score for pairs of DBpedia entities. For example, the DBpedia entity `Dishwasher` has a cosine similarity of .803 to the entity `Kitchen`, but only .279 with `Classroom`, suggesting that the prototypical location for a generic dishwasher is the kitchen rather than a classroom. Since cosine similarity is a graded value on a scale from -1 to 1, we can generate, for a given object, a ranking of candidate locations, e.g. the rooms of a house. Table 3 shows a sample of object-location pairs of DBpedia labels, ordered by the cosine similarity of their respective vectors in NASARI. Prototypical locations for the objects show up at the top of the list as expected, indicating a relationship between the semantic relatedness expressed by the cosine similarity of vector representations and the actual locative relation of entities. We refer to this model as `NASARI/Cosine`.

3.3. Ranking by a Trained Scoring Function

In the previous sections, we presented models of semantic relatedness for the extraction of relations. The employed cosine similarity function of these models is relation-agnostic, that is, it only measures whether there is a relation between two entities but not which relation in particular. The question that naturally arises is: *Instead of using a single model that is agnostic to the relation, can we train a separate model for each relation in order to improve the extraction performance?* In this section we try to answer this question by introducing a new model, based on supervised learning.

Table 3
Locations for a sample object, extracted by computing cosine similarity on NASARI vectors.

Object	Location	Cos. Similarity
Dishwasher	Kitchen	.803
	Air_shower_(room)	.788
	Utility_room	.763
	Bathroom	.758
	Furnace_room	.749
Paper_towel	Air_shower_(room)	.671
	Public_toilet	.634
	Bathroom	.632
	Mizuya	.597
	Kitchen	.589
Sump_pump	Furnace_room	.699
	Air_shower_(room)	.683
	Basement	.680
	Mechanical_room	.676

To extend the proposed approach to any kind of relation we modify the model presented in Section 3.1 by introducing a parameterized scoring function. This scoring function replaces the cosine similarity which was previously employed to score pairs of entities (e.g. Object-Location). By tuning the parameters of this new scoring function in a data-driven way, we are able to predict scores with respect to arbitrary relations.

We define the new scoring function $f(s,r,o)$ as a bilinear form:

$$f(s,r,o) = tanh(\vec{v}_s^{\top} \boldsymbol{M}_r \vec{v}_o + b_r) \tag{1}$$

where $\vec{v}_s, \vec{v}_o \in \boldsymbol{V} \subseteq \mathbb{R}^d$ are the corresponding embedding vectors for the subject and object entities s and o, respectively, b_r is a bias term, and $\boldsymbol{M}_r \in \mathbb{R}^{d \times d}$ is the scoring matrix corresponding to the relation r. Our scoring function is closely related to the ones proposed by Jenatton et al. [30] as well as Yang et al. [71], however, we make use of the *tanh* activation function to map the scores to the interval $(-1,1)$. In part, this relates to the Neural Tensor Network proposed by Socher et al. [60]. By initializing $\boldsymbol{M}_r$ as the identity matrix and b_r with 0, the inner term of the scoring function corresponds initially to the dot product of $\vec{v}_s$ and $\vec{v}_o$ which is closely related to the originally employed cosine similarity.

In order to learn the parameters $\boldsymbol{M}_r$ and b_r of the scoring function, we follow a procedure related to

Noise Contrastive Estimation [50] and Negative Sampling [44] which is also used in the training of the skip-gram embeddings. This method uses "positive" and "negative" triples, $\mathcal{T}^+_{train}$ and $\mathcal{T}^-_{train}$, to iteratively adapt the parameters. The positive triples $\mathcal{T}^+_{train}$ are triples that truly express the respective relation. In our case, these triples are obtained by crowdsourcing and leveraging other resources (see Section 4). Given these positive triples, the set of corrupted negative triples $\mathcal{T}^-_{train}$ is generated in the following way: We generate negative triples (s', r, o) and (s, r, o') for each positive triple $(s, r, o) \in \mathcal{T}^+$ by selecting negative subject and object entities s' and o' randomly from the set of all possible subjects and objects, respectively. The exact number of negative triples that we generate per positive triple is a hyper-parameter of the model which we set to 10 triples[8] for all our experiments.

The training of the scoring function is framed as a classification where we try to assign scores of 1 to all positive triples and scores of -1 to (randomly generated) negative triples. We employ the mean squared error (MSE) as the training objective:

$$\mathcal{L} = \frac{1}{N}\Bigg(\sum_{(s,r,o)\in\mathcal{T}^+_{train}} (1 - f(s,r,o))^2 + \sum_{(s,r,o)\in\mathcal{T}^-_{train}} (-1 - f(s,r,o))^2 \Bigg) \tag{2}$$

where $N = |\mathcal{T}^+_{train}| + |\mathcal{T}^-_{train}|$ is the size of the complete training set. During training, we keep the embedding vectors $\boldsymbol{V}$ fixed and only consider $\boldsymbol{M}_r$ and b_r as trainable parameters to measure the effect of the scoring function in isolation. Presumably, this allows for a better generalization to previously unseen entities.

Due to the moderate size of our training data, we regularize our model by applying Dropout [62] to the embedding vectors of the head and tail entity. We set the dropout fraction to 0.1, thus only dropping a small portion of the 100 dimensional input vectors.

The supervised model differs from the unsupervised approaches in that the scoring function is tuned to a particular relation, e.g. the `locatedAt` relation from Section 4. In the following, we denote this model as `SkipGram/Supervised`.

[8] 5 triples (s', r, o) where we corrupt the subject entity and 5 triples (s, r, o') where the object entity is replaced.

4. Datasets

The following section introduces the datasets that we use for this work. We consider three types of datasets: i) a crowdsourced set of triples expressing the `locatedAt` relation with human judgments, ii) a semi-automatically extracted set of triples expressing the `locatedAt` relation, and iii) a semi-automatically extracted set of `usedFor` triples.

4.1. Crowdsourcing of Object-Location Rankings

In order to acquire valid pairs for the `locatedAt` relation we rely on a crowdsourcing approach. In particular, given a certain object, we used crowdsourcing to collect judgments about the likelihood to find this object at a set of predefined locations.

To select the objects and locations for this experiment, every DBpedia entity that falls under the category `Domestic_implements`, or under one of the `narrower` categories than `Domestic_implements` according to SKOS[9], is considered an object. The SPARQL query is given as:

```
select distinct ?object where {
 {
  ?object
    <http://purl.org/dc/terms/subject>
    dbc:Domestic_implements
 } UNION {
  ?object
    <http://purl.org/dc/terms/subject>
    ?category .
  ?category
    <http://www.w3.org/2004/02/skos/core#broader>
    dbc:Domestic_implements .
 }
}
```

Every DBpedia entity that falls under the category `Rooms` is considered a location. The respective query is:

```
select distinct ?room where {
?room
  <http://purl.org/dc/terms/subject>
  dbc:Rooms
}
```

These steps result in 336 objects and 199 locations (as of September 2016). To select suitable pairs expressing the `locatedAt` relation for the creation of the gold standard, we filter out odd or uncommon examples of objects or locations like `Ghodiyu` or `Fainting_room`. We do this by ordering the objects by the number of incoming links to their respective Wikipedia page[10] in descending order and select the

[9] Simple Knowledge Organization System: `https://www.w3.org/2004/02/skos/`

[10] We use the URI counts extracted from the parsing of Wikipedia with the DBpedia Spotlight tool for entity linking [17].

100 top ranking objects for our gold standard. We proceed analogously for the locations, selecting 20 common locations and thus obtain 2,000 object-location pairs in total.

In order to collect the judgments, we set up a crowdsourcing experiment on the CrowdFlower platform[11]. For each of the 2,000 object-location pairs, contributors were asked to rate the likelihood of the object to be in that location on a four-point scale:

- **-2** (**unexpected**): finding the object in the room would cause surprise, e.g. it is unexpected to find a bathtub in a cafeteria.
- **-1** (**unusual**): finding the object in the room would be odd, the object feels out of place, e.g. it is unusual to find a mug in a garage.
- **1** (**plausible**): finding the object in the room would not cause any surprise, it is seen as a normal occurrence, e.g. it is plausible to find a funnel in a dining room.
- **2** (**usual**): the room is the place where the object is typically found, e.g, the kitchen is the usual place to find a spoon.

Contributors were shown ten examples per page, instructions, a short description of the entities (the first sentence from the Wikipedia abstract), a picture (from Wikimedia Commons, when available[12]), and the list of possible answers as labeled radio buttons.

After running the crowdsourcing experiment for a few hours, we collected 12,767 valid judgments, whereas 455 judgments were deemed "untrusted" by CrowdFlower's quality filtering system. The quality control was based on 57 test questions that we provided and a required minimum accuracy of 60% on these questions for a contributor to be considered trustworthy. In total, 440 contributors participated in the experiment.

The pairs received on average 8.59 judgments. Most of the pairs received at least 5 separate judgments, with some outliers collecting more than one hundred judgments each. The average agreement, i.e. the percentage of contributors that answered the most common answer for a given question, is 64.74%. The judgments are skewed towards the negative end of the spectrum, as expected, with 37% pairs rated unexpected, 30% unusual, 24% plausible and 9% usual. The cost of the experiment was 86 USD.

[11] `http://www.crowdflower.com/`

[12] Pictures were available for 94 out of 100 objects.

To use this manually labeled data in later experiments, we normalize, filter and rearrange the scored pairs and obtain three gold standard datasets:

For the first gold standard dataset, we reduce multiple human judgments for each Object-Location pair to a single score by assigning the average of the numeric values. For instance, if the pair (`Wallet, Ballroom`) has been rated -2 (unexpected) six times, -1 (unusual) three times, and never 1 (plausible) or 2 (usual), its score will be about -1.6, indicating that a `Wallet` is not very likely to be found in a `Ballroom`. For each object, we then produce a ranking of all 20 locations by ordering them by their averaged score for the given object. We refer to this dataset of human-labeled rankings as *locatedAt-Human-rankings*.

The second and third gold standard datasets are produced as follows: The contributors' answers are aggregated using relative majority, that is, each object-location pair has exactly one judgment assigned to it, corresponding to the most popular judgment among all the contributors that answered that question. We extract two sets of relations from this dataset to be used as a gold standard for experimental tests: one list of the 156 pairs rated 2 (*usual*) by the majority of contributors, and a larger list of the 496 pairs rated either 1 (*plausible*) or 2 (*usual*). The aggregated judgments in the gold standard have a confidence score assigned to them by CrowdFlower, based on a measure of inter-rater agreement. Pairs that score low on this confidence measure ($\leqslant 0.5$) were filtered out, leaving 118 and 496 pairs, respectively. We refer to these two gold standard sets as *locatedAt-usual* and *locatedAt-usual/plausible*.

4.2. Semi-Supervised Extraction of Object-Location Triples

The SUN database [70] is a large-scale resource for computer vision and object recognition in images. It comprises 131,067 single images, each of them annotated with a label for the type of scene, and labels for each object identified in the scene. The images are annotated with 908 categories based on the type of scene (bedroom, garden, airway, ...). Moreover, 313,884 objects were recognized and annotated with one out of 4,479 category labels.

Despite its original goal of providing high-quality data for training computer vision models, the SUN project generated a wealth of semantic knowledge that is independent from the vision tasks. In particular, the labels are effectively semantic categories of entities

Table 4
Most frequent pairs of object-scene in the SUN database.

Frequency	Object	Scene
1041	wall	b/bedroom
1011	bed	b/bedroom
949	floor	b/bedroom
663	desk_lamp	b/bedroom
650	night_table	b/bedroom
575	ceiling	b/bedroom
566	window	b/bedroom
473	pillow	b/bedroom
463	wall	b/bathroom
460	curtain	b/bedroom
406	painting	b/bedroom
396	floor	b/bathroom
393	cushion	b/bedroom
380	wall	k/kitchen
370	wall	d/dining_room
364	chair	d/dining_room
355	table	d/dining_room
351	floor	d/dining_room
349	cabinet	k/kitchen
344	sky	s/skyscraper

such as objects and locations (scenes, using the lexical conventions of the SUN database).

Objects are observed at particular scenes, and this relational information is retained in the database. In total, we extracted 31,407 object-scene pairs from SUN, together with the number of occurrences of each pair. The twenty most occurring pairs are shown in Table 4.

According to its documentation, the labels of the SUN database are lemmas from WordNet. However, they are not disambiguated and thus they could refer to any meaning of the lemma. Most importantly for our goals, the labels in their current state are not directly linked to any LOD resource. Faced with the problem of mapping the SUN database completely to a resource like DBpedia, we adopted a safe strategy for the sake of the gold standard creation. We took all the object and scene labels from the SUN pairs for which a resource in DBpedia with matching label exists. In order to limit the noise and obtain a dataset of "typical" location relations, we also removed those pairs that only occur once in the SUN database. This process resulted in 2,961 pairs of entities. We manually checked them and corrected 118 object labels and 44 location labels. In some cases the correct label was already present, so we eliminated the duplicates resulting in a new dataset of 2,935 object-location pairs[13]. The collected triples are used in Sections 5.1 and 5.2 as training data. We refer to this dataset as *locatedAt-Extracted-triples*.

4.3. Semi-Supervised Extraction of Object-Action Triples

While the methods we propose for relation extractions are by design independent of the particular relations they are applied to, we have focused most of our experimental effort towards one kind of relation between objects and locations, namely the typical location where given objects are found. As a first step to assess the generalizability of our approaches to other kinds of relations, we created an alternative dataset revolving around a relation with the same domain as the location relation, i.e., objects, but a very different range, that is, actions. The relation under consideration will be referred to in the rest of the article as `usedFor`, for example the predicate `usedFor(soap, bath)` states that the soap is used for (or, during, in the process of) taking a bath.

We built a dataset of object-action pairs in a `usedFor` relation starting from ConceptNet 5 [39], a large semantic network of automatically collected commonsense facts (see also Section 2). From the entire ConceptNet, we extracted 46,522 links labeled `usedFor`. Although ConceptNet is partly linked to LOD resources, we found the coverage of such linking to be quite low, especially with respect to non-named entities such as objects. Therefore, we devised a strategy to link as many of the labels involved in `usedFor` relations to DBpedia, without risking to compromise the accuracy of such linking. The strategy is quite simple and it starts from the observation of the data: for the first argument of the relation, we search DBpedia for an entity whose label matches the ConceptNet labels. For the second argument, we search DBpedia for an entity label that matches the gerund form of the ConceptNet label, e.g. *Bath→Bathing*. We perform this step because we noticed how actions are usually referred to with nouns in ConceptNet, but with verbs in the gerund form in DBpedia. We used the morphology generation tool for English *morphg* [46] to generate the correct gerund forms also for irregular verbs. The application of this linking strategy resulted in a dataset of 1,674 pairs of DBpedia entities. Table 5 shows a few examples of pairs in the dataset.

[13] Of all extracted triples, 24 objects and 12 locations were also among the objects and locations of the crowdsourced dataset.

Table 5
Examples of DBpedia entities in a `usedFor` relation, according to ConceptNet and our DBpedia linking strategy.

Object	Action
Machine	Drying
Dictionary	Looking
Ban	Saving
Cake	Jumping
Moon	Lighting
Tourniquet	Saving
Dollar	Saving
Rainbow	Finding
Fast_food_restaurant	Meeting
Clipboard	Keeping

To use this data as training and test data for the proposed models, we randomly divide the complete set of positive (Object, usedFor, Action) triples in a training portion (527 triples) and a test portion (58 triples). We combine each object entity in the test portion with each action entity to generate a complete test set, comprised of positive and negative triples[14]. To account for variations in the performance due to this random partitioning, we repeat each experiment 100 times and report the averaged results in the experiments in Section 5.3. The average size of the test set is ≈ 2059. We refer to this dataset as *usedFor-Extracted-triples*.

5. Evaluation

This section presents the evaluation of the proposed framework for relation extraction (Sections 3.1, 3.2 and 3.3). We apply our models to the data described in Section 4, consisting of sets of (Object, locatedAt, Location) and (Object, usedFor, Action) triples. These experiments verify the feasibility of our approach for the population of a knowledge base of manipulation relevant data.

We start our experiments by evaluating how well the produced rankings of (Object, locatedAt, Location) triples match the ground truth rankings obtained from human judgments. For this, we i) present the evaluations for the unsupervised methods SkipGram/Cosine and NASARI/Cosine (Section 5.1.1), ii) show the performance of combinations thereof (Section 5.1.2) and iii) evaluate the newly proposed SkipGram/Supervised method (Section 5.1.3).

[14] We filter out all generated triples that are falsely labeled as negative in this process.

The second part of our experiments evaluates how well each proposed method performs in extracting a knowledge base. The evaluation is performed for (Object, locatedAt, Location) and (Object, usedFor, Action) triples (Sections 5.2 and 5.3, respectively).

5.1. Ranking Evaluation

With the proposed methods from previous sections, we are able to produce a ranking of e.g. locations for a given object that expresses how prototypical the location is for that object. To test the validity of our methods, we compare their output against the gold standard rankings *locatedAt-Human-rankings* that we obtained from the crowdsourced pairs (see Section 4.1).

As a first evaluation, we investigate how well the unsupervised baseline methods perform in creating object-location rankings. Secondly, we show how to improve these results by combining different approaches. Thirdly, we evaluate the supervised model in comparison to our baselines.

5.1.1. Unsupervised Object-Location Ranking Evaluation

Apart from the NASARI-based method (Section 3.2) and the skip-gram-based method (Section 3.1) we employ two simple baselines for comparison: For the *location frequency* baseline, the object-location pairs are ranked according to the frequency of the location. The ranking is thus the same for each object, since the score of a pair is only computed based on the location. This method makes sense in absence of any further information on the object: e.g. a robot tasked to find an unknown object should inspect "common" rooms such as a kitchen or a studio first, rather than "uncommon" rooms such as a pantry.

The second baseline, the *link frequency*, is based on counting how often every object appears on the Wikipedia page of every location and vice versa. A ranking is produced based on these counts. An issue with this baseline is that the collected counts could be sparse, i.e., most object-location pairs have a count of 0, thus sometimes producing no value for the ranking for an object. This is the case for rather "unusual" objects and locations.

For each object in the dataset, we compare the location ranking produced by our algorithms to the crowdsourced gold standard ranking and compute two metrics: the *Normalized Discounted Cumulative Gain* (NDCG) and the *Precision at k* (Precision@k or P@k).

Table 6
Average Precision@k for $k = 1$ and $k = 3$ and average NDCG of the produced rankings against the gold standard rankings.

Method	NDCG	P@1	P@3
Location frequency baseline	.851	.000	.008
Link frequency baseline	.875	.280	.260
NASARI/Cosine	.903	**.390**	.380
SkipGram/Cosine	**.912**	.350	**.400**

The NDCG is a measure of rank correlation used in information retrieval that gives more weight to the results at the top of the list than at its bottom. It is defined as follows:

$$NDCG(R) = \frac{DCG(R)}{DCG(R^*)}$$

$$DCG(R) = R_1 + \sum_{i=2}^{|R|} \frac{R_i}{log_2(i+1)}$$

where R is the produced ranking, R_i is the true relevance of the element at position i and R^* is the ideal ranking of the elements in R. R^* can be obtained by sorting the elements by their true relevance scores. This choice of evaluation metric follows from the idea that it is more important to accurately predict which locations are likely for a given object than to decide which are unlikely candidates.

While the NDCG measure gives a complete account of the quality of the produced rankings, it is not easy to interpret apart from comparisons of different outputs. To gain a better insight into our results, we provide an alternative evaluation, the *Precision@k*. The Precision@k measures the number of locations among the first k positions of the produced rankings that are also among the top-k locations in the gold standard ranking. It follows that, with $k = 1$, precision at 1 is 1 if the top returned location is the top location in the gold standard, and 0 otherwise. We compute the average of Precision@k for $k = 1$ and $k = 3$ across all the objects.

Table 6 shows the average NDCG and Precision@k across all objects: methods NASARI/Cosine (Section 3.2) and SkipGram/Cosine (Section 3.1), plus the two baselines introduced above.

Both our methods that are based on semantic relatedness outperform the simple baselines with respect to the gold standard rankings. The location frequency baseline performs very poorly, due to an idiosyncrasy in the frequency data, that is, the most "frequent" location in the dataset is *Aisle*. This behavior reflects the difficulty in evaluating this task using only automatic metrics, since automatically extracted scores and rankings may not correspond to common sense judgment.

The NASARI-based similarities outperform the skip-gram-based method when it comes to guessing the most likely location for an object (Precision@1), as opposed to the better performance of SkipGram/Cosine in terms of Precision@3 and rank correlation.

We explored the results and found that for 19 objects out of 100, NASARI/Cosine correctly guesses the top ranking location where SkipGram/Cosine fails, while the opposite happens 15 out of 100 times. We also found that the NASARI-based method has a lower coverage than the skip-gram method, due to the coverage of the original resource (NASARI), where not every entity in DBpedia is assigned a vector[15]. The skip-gram-based method also suffers from this problem, however, only for very rare or uncommon objects and locations (as `Triclinium` or `Jamonera`). These findings suggest that the two methods could have different strengths and weaknesses. In the following section we show two strategies to combine them.

5.1.2. *Hybrid Methods: Fallback Pipeline and Linear Combination*

The results from the previous sections highlight that the performance of our two main methods may differ qualitatively. In an effort to overcome the coverage issue of NASARI/Cosine, and at the same time experiment with hybrid methods to extract location relations, we devised two simple ways of combining the SkipGram/Cosine and NASARI/Cosine methods. The first method is based on a fallback strategy: given an object, we consider the pair similarity of the object to the top ranking location according to NASARI/Cosine as a measure of confidence. If the top ranked location among the NASARI/Cosine ranking is exceeding a certain threshold, we consider the ranking returned by NASARI/Cosine as reliable. Otherwise, if the similarity is below the threshold, we deem the result unreliable and we adopt the ranking returned by SkipGram/Cosine instead. The second method produces object-location similarity scores by linear combination of the NASARI and skip-gram similarities. The similarity score for the generic pair s, o is thus given by:

$$\begin{aligned} sim_\alpha(s,o) = & \quad \alpha \cdot sim_{NASARI}(s,o) \\ & + \quad (1-\alpha) \cdot sim_{SkipGram}(s,o), \end{aligned} \tag{3}$$

[15] Objects like `Backpack` and `Comb`, and locations like `Loft` are all missing.

Table 7
Rank correlation and precision at k for the method based on fallback strategy.

Method	NDCG	P@1	P@3
Fallback strategy (threshold=.4)	.907	**.410**	.393
Fallback strategy (threshold=.5)	.906	.400	.393
Fallback strategy (threshold=.6)	.908	**.410**	.406
Fallback strategy (threshold=.7)	.909	.370	.396
Fallback strategy (threshold=.8)	.911	.360	.403
Linear combination (α=.0)	.912	.350	.400
Linear combination (α=.2)	.911	.380	.407
Linear combination (α=.4)	**.913**	.400	**.423**
Linear combination (α=.6)	.911	.390	.417
Linear combination (α=.8)	.910	.390	.410
Linear combination (α=1.0)	.903	.390	.380

where parameter α controls the weight of one method with respect to the other.

Table 7 shows the obtained results, with varying values of the parameters *threshold* and α. While the NDCG is only moderately affected, both Precision@1 and Precision@3 show an increase in performance with Precision@3 showing the highest score of all investigated methods.

5.1.3. Supervised Object-Location Ranking

In the previous experiments, we investigated how well our (unsupervised) baseline methods perform when extracting the `locatedAt` relation. In the following, we compare the earlier results to the performance of a scoring function trained in a supervised fashion. For this experiment we train the scoring function in Eq. (1) to extract the `locatedAt` relation between objects and locations. The underlying embeddings $\boldsymbol{V}$ on which the scoring function computes its scores are fixed to the skip-gram embeddings $\boldsymbol{V}_{sg}$ (see Section 3.1). We train the supervised method on the semi-automatically extracted triples *locatedAt-Extracted-triples* described in Section 4.2. These triples act as the positive triples $\mathcal{T}^{+}_{train}$ in the training procedure, from which we also generate the negative examples $\mathcal{T}^{-}_{train}$ following the procedure in Section 3.3. As described in Section 3.3, we train the model by generating 10 negative triples per positive triple and minimizing the mean squared error from Eq. (2). We initialize $\mathbf{M}_r$ with the identity matrix, b_r with 0, and train the model parameter using stochastic gradient descent (SGD) using a learning rate of 0.001. SGD is performed in mini batches of size 100 with 300 epochs of training. The training procedure is realized with *Keras* [13].

Table 8
Average precision at k for $k = 1$ and $k = 3$ and average NDCG of the produced rankings against the crowdsourced gold standard rankings. *SkipGram/Supervised* denotes the supervised model based on skip-gram embeddings trained for the `locatedAt` relation.

Method	NDCG	P@1	P@3
Location frequency baseline	.851	.000	.008
Link frequency baseline	.875	.280	.260
NASARI/Cosine	.903	.390	.380
SkipGram/Cosine	.912	.350	.400
Linear combination (α=.4)	**.913**	.400	**.423**
SkipGram/Supervised	.908	**.454**	.387

As before, we test the model on the human-rated set of objects and locations *locatedAt-Human-rankings* described in Section 4.1 and produce a ranking of locations for each object. Table 8 shows the performance of the extended model (SkipGram/Supervised) in comparison to the previous approaches.

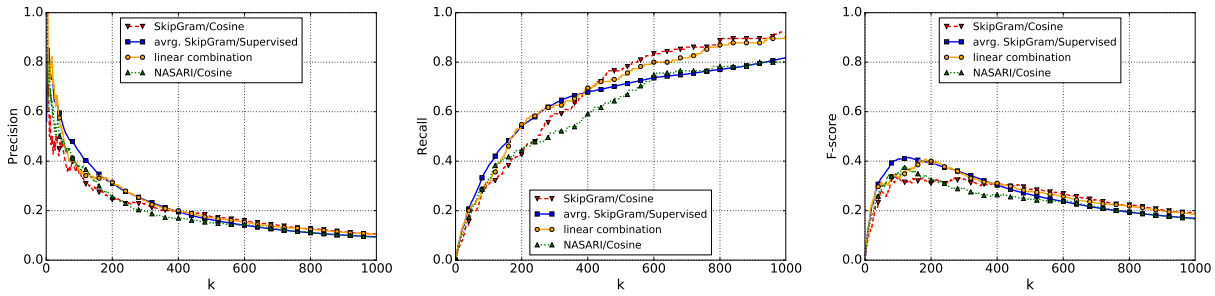
Overall, we can observe mixed results. All of our proposed models (supervised and unsupervised) improve upon the baseline methods with respect to all evaluation metrics. Compared to the SkipGram/Cosine model, the SkipGram/Supervised model decreases slightly in performance with respect to the NDCG and more so for the Precision@3 score. Most striking, however, is the increase in Precision@1 of SkipGram/-Supervised, showing a relative improvement of 30% to the SkipGram/Cosine model and constituting the highest overall Precision@1 score by a large margin. However, the linear combination (α=.4) still scores higher with respect to Precision@3 and NDCG.

While the presented results do not point to a clear preference for one particular model, Section 5.2 will investigate the above methods more closely in the context of the generation of a knowledge base.

5.2. Retrieval Evaluation

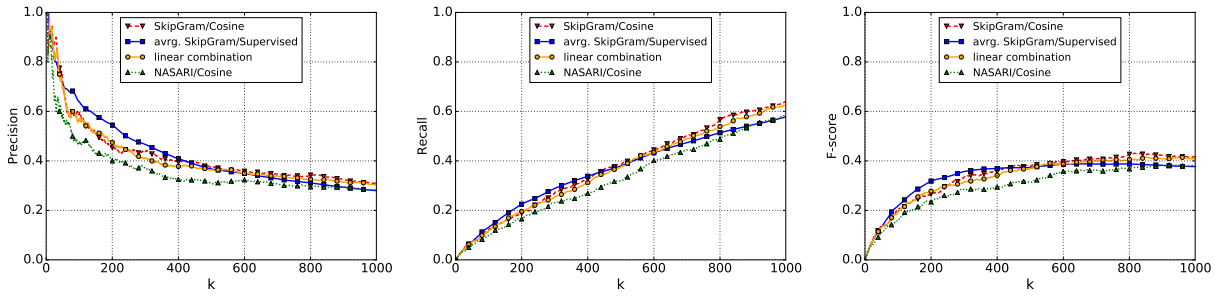
In the previous section, we tested how the proposed methods perform in determining a ranking of locations given an object. For the purpose of evaluation, the tests have been conducted on a closed set of entities. In this section we return to the original motivation of this work, that is, to collect manipulation-relevant information about objects in an automated fashion in the form of a knowledge base.

All the methods introduced in this work are based on some scoring function of triples expressed as a real number in the range [-1,1] and thus interpretable as a sort of confidence score relative to the target relation.



(a) Precision (b) Recall (c) F-score

Fig. 1. Evaluation on automatically created knowledge bases (“usual” locations).



(a) Precision (b) Recall (c) F-score

Fig. 2. Evaluation on automatically created knowledge bases (“plausible” and “usual” locations).

Therefore, by imposing a threshold on the similarity scores and selecting only the object-location pairs that score above said threshold, we can extract a high-confidence set of object-location relations to build a new knowledge base from scratch. Moreover, by using different values for the threshold, we are able to control the quality and the coverage of the produced relations. We test this approach on:

- the *locatedAt-usual* and *locatedAt-usual/plausible* datasets (Section 4.1) for the `locatedAt` relation between objects and locations, and
- the *usedFor-Extracted-triples* dataset (Section 4.3) for the `usedFor` relation between objects and actions.

We introduce the `usedFor` relation in order to assess the generalizability of our supervised scoring function.

In general, we extract a knowledge base of triples by scoring each possible candidate triple, thus producing an overall ranking. We then select the top k triples from the ranking, with k being a parameter. This gives us the triples that are considered the most prototypical. We evaluate the retrieved set in terms of Precision, Recall and F-score against the gold standard sets with varying values of k. Here, the precision is the fraction of correctly retrieved triples in the set of all retrieved triples, while the recall is the fraction of retrieved triples that also occur in the gold standard set. The F-score is the harmonic mean of precision and recall:

$$Precision = \frac{|\mathcal{G} \cap \mathcal{R}_k|}{|\mathcal{R}_k|}$$

$$Recall = \frac{|\mathcal{G} \cap \mathcal{R}_k|}{|\mathcal{G}|}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

with $\mathcal{G}$ denoting the set of gold standard triples and $\mathcal{R}_k$ the set of retrieved triples up to rank k.

For the `locatedAt` relation, we also add to the comparison the results of the hybrid, linear combination method from Section 5.1.2, with the best performing parameters in terms of Precision@1, namely the linear combination with $\alpha = 0.4$.

Figures 1 and 2 show the evaluation of the four methods evaluated against the two aggregated gold standard datasets for the `locatedAt` relation described above. Figures 1c and 2c, in particular, show F-score plots for a direct comparison of the performance. The SkipGram/Supervised model achieves the highest F-score on the *locatedAt-usual* dataset, peaking at $k = 132$ with an F-score of 0.415. The SkipGram/Cosine model and the linear combination outperform both the NASARI/Cosine and the SkipGram/-Supervised in terms of recall, especially for higher k. This also holds for the *locatedAt-usual/plausible* dataset. Here, the SkipGram/Supervised model stands out by achieving high precision values for small values of k. Overall, SkipGram/Supervised performs better for small k (50 – 400) whereas SkipGram/Cosine and the linear combination obtain better results with increasing k. This seems to be in line with the results from previous experiments in Table 8 that show a high Precision@1 for the SkipGram/Supervised model but higher scores for SkipGram/Cosine and the linear combination in terms of Precision@3.

5.3. Evaluation of Object-Action pairs extraction

One of the reasons to introduce a novel technique for relation extraction based on a supervised statistical method, as stated previously, is to be able to scale the extraction across different types of relations. To test the validity of this statement, we apply the same evaluation procedure introduced in the previous part of this section to the `usedFor` relation. For the training and evaluation sets we use the dataset *usedFor-Extracted-triples* comprising of semi-automatically extracted triples from ConceptNet (Section 4.3).

Figure 3 displays precision, recall and F-score for retrieving the top k results. The results are averaged scores over 100 experiments to account for variations in performance due to the random partitioning in training and evaluation triples and the generation of negative samples. The standard deviation for precision, recall and F-score for all k is visualized along the mean scores.

The supervised model achieves on average a maximum F-score of about 0.465 when extracting 70 triples. This is comparable to the achieved F-scores when training the scoring function for the `locatedAt` relation. To give an insight into the produced false positives, Table 9 shows the top 30 extracted triples for the `usedFor` relation of one trained instance of the supervised model.

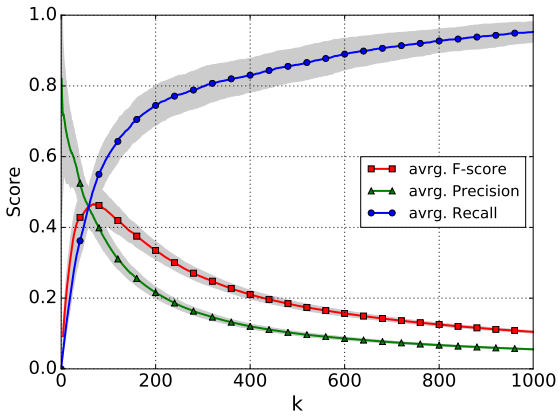


Fig. 3. Evaluation of knowledge base generation for the `usedFor` relation between objects and actions. Precision, Recall and F-score are given with respect to extracting the to k scored triples.

6. Building a Knowledge Base of Object Locations

Given these results, we can aim for a high-confidence knowledge base by selecting the threshold on object-location similarity scores that produces a reasonably high precision knowledge base in the evaluation. For instance, the knowledge base made by the top 50 object-location pairs extracted with the linear combination method ($\alpha = 0.4$) has 0.52 precision and 0.22 recall on the *locatedAt-usual* gold standard (0.70 and 0.07 respectively on the *locatedAt-usual/plausible* set, see Figures 1a and 2a). The similarity scores in this knowledge base range from 0.570 to 0.866. Following the same methodology that we used to construct the gold standard set of objects and locations (Section 4.1), we extract all the 336 `Domestic_implements` and 199 `Rooms` from DBpedia, for a total of 66,864 object-location pairs. Selecting only the pairs whose similarity score is higher than 0.570, according to the linear combination method, yields 931 high confidence location relations. Of these, only 52 were in the gold standard set of pairs (45 were rated "usual" or "plausible" locations), while the remaining 879 are new, such as (`Trivet`, `Kitchen`), (`Flight_bag`, `Airport_lounge`) or (`Soap_dispenser`, `Unisex_public_toilet`). The distribution of objects across locations has an arithmetic mean of 8.9 objects per location and standard deviation 11.0. `Kitchen` is the most represented

Table 9
A list of the top 30 extracted triples for the usedFor relation. The gray highlighted rows mark the entity pairs that are part of the gold standard dataset (Section 4.3).

Score	Object	Action
1.00000	Snack	Snacking
0.99896	Snack	Eating
0.99831	Curry	Seasoning
0.99773	Drink	Drinking
0.98675	Garlic	Seasoning
0.98165	Oatmeal	Snacking
0.98120	Food	Eating
0.96440	Pistol	Shooting
0.95218	Drink	Snacking
0.94988	Bagel	Snacking
0.94926	Wheat	Snacking
0.93778	Laser	Printing
0.92760	Food	Snacking
0.91946	Typewriter	Typing
0.91932	Oatmeal	Eating
0.91310	Wok	Cooking
0.89493	Camera	Shooting
0.85415	Coconut	Seasoning
0.85091	Stove	Frying
0.85039	Oatmeal	Seasoning
0.84038	Bagel	Eating
0.83405	Cash	Gambling
0.81985	Oatmeal	Baking
0.80975	Lantern	Lighting
0.80129	Calculator	Typing
0.78279	Laser	Shooting
0.77411	Camera	Recording
0.75712	Book	Writing
0.72924	Stove	Cooking
0.72280	Coconut	Snacking

location with 89 relations, while 15 out of 107 locations are associated with one single object.[16]

The knowledge base created with this method is the result of one among many possible configurations of a number of methods and parameters. In particular, the creator of a knowledge base involving the extraction of relations is given the choice to prefer precision over recall, or vice-versa. This is done, in our method, by adjusting the threshold on the similarity scores. Employing different algorithms for the computation of the actual similarities (word embeddings vs. entity vectors, supervised vs. unsupervised models) is also expected to result in different knowledge bases. A qualitative assessment of such impact is left for future work.

[16]The full automatically created knowledge base and used resources are available at https://project.inria.fr/aloof/data/.

7. Conclusion and Future Work

We have presented a framework for extracting manipulation-relevant knowledge about objects in the form of (binary) relations. The framework relies on a ranking measure that, given an object, ranks all entities that potentially stand in the relation in question to the given object. We rely on a representational approach that exploits distributional spaces to embed entities into low-dimensional spaces in which the ranking measure can be evaluated. We have presented results on two relations: the relation between an object and its prototypical location (locatedAt) as well as the relation between an object and one of its intended uses (usedFor).

We have shown that both an approach relying on standard word embeddings computed by a skip-gram model as well as an approach using embeddings computed for disambiguated concepts rather than lemmas perform very well compared to two rather naive baselines. Both approaches were presented already in previous work. As main contribution of this paper, we have presented a supervised approach based on a neural network that, instead of using the cosine similarity as measure of semantic relatedness, uses positive and negative examples to train a scoring function in a supervised fashion. In contrast to the other two unsupervised approaches, the latter learns a model that is specific for a particular relation while the other two approaches implement a general notion of semantic relatedness in distributional space.

We have shown that the improvements of the supervised model are not always clear compared to the two unsupervised approaches. This might be attributable to the fact that the types of both relations (usedFor and locatedAt) are specific enough to predict the relation in question. Whether the unsupervised approach would generalize to relations with a less specific type signature remains to be seen.

As an avenue for future work, the generalizability of the proposed methods to a wider set of relations can be considered. In the context of manipulation-relevant knowledge for a robotic system, other interesting properties of an object include its prototypical size, weight, texture, and fragility. Additionally, we see possibilities to address relations as can be found in ConceptNet

5 [61] such as MadeOf, Causes, CausesDesire, CapableOf, and more that all help a robot to interact with humans and objects in its environment.

We also plan to employ *retrofitting* [22] to enrich our pretrained word embeddings with concept knowledge from a semantic network such as ConceptNet or WordNet [45] in a post-processing step. With this technique, we might be able to combine the benefits of the concept-level and word-level semantics in a more sophisticated way to bootstrap the creation of an object-location knowledge base. We believe that this method is a more appropriate tool than the simple linear combination of scores. By specializing our skip-gram embeddings for relatedness instead of similarity [31] even better results could be achieved.

In the presented work, we used the frequency of entity mentions in Wikipedia as a measure of commonality to drive the creation of a gold standard set for evaluation. This information, or equivalent measures, could be integrated directly into our relation extraction framework, for example in the form of a weighting scheme or hand-crafted features, to improve its prediction accuracy.

Acknowledgments The authors wish to thank the anonymous reviewers of EKAW, who provided useful feedback to make this extended version of the paper. The work in this paper is partially funded by the ALOOF project (CHIST-ERA program) and by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

References

[1] Ken Barker, Bhalchandra Agashe, Shaw Yi Chaw, James Fan, Noah S. Friedland, Michael Robert Glass, Jerry R. Hobbs, Eduard H. Hovy, David J. Israel, Doo Soon Kim, Rutu Mulkar-Mehta, Sourabh Patwardhan, Bruce W. Porter, Dan Tecuci, and Peter Z. Yeh. Learning by reading: A prototype system, performance baseline and lessons learned. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 280–286, 2007.

[2] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247, June 2014. DOI: 10.3115/v1/P14-1023.

[3] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] Lawrence W. Barsalou. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1281–1289, 2009. DOI: 10.1098/rstb.2008.0319.

[5] Valerio Basile, Soufian Jebbara, Elena Cabrio, and Philipp Cimiano. Populating a Knowledge Base with Object-Location Relations Using Distributional Semantics. In *20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016)*, pages 34 – 50, Bologna, Italy, November 2016. DOI: 10.1007/978-3-319-49004-5_3.

[6] Sebastian Blohm and Philipp Cimiano. Using the web to reduce data sparseness in pattern-based information extraction. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, pages 18–29, 2007. DOI: 10.1007/978-3-540-74976-9_6.

[7] Sebastian Blohm, Philipp Cimiano, and Egon Stemle. Harvesting relations from the web -quantifiying the impact of filtering functions. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pages 1316–1323. Association for the Advancement of Artificial Intelligence (AAAI), Juli 2007.

[8] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM. DOI: 10.1145/1376616.1376746.

[9] Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER 2012, June 18-21, 2012, Québec City, Canada*, pages 481–495. 2012. DOI: 10.1007/978-3-319-00065-7_33.

[10] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc., 2013.

[11] Razvan C. Bunescu and Raymond J. Mooney. Subsequence kernels for relation extraction. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 171–178, Cambridge, MA, USA, 2005. MIT Press.

[12] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. NASARI: A novel approach to a semantically-aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 567–577, 2015. DOI: 10.3115/v1/N15-1059.

[13] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[14] Philipp Cimiano and Johanna Wenderoth. Automatically Learning Qualia Structures from the Web. In Timothy Baldwin, Anna Korhonen, and Aline Villavicencio, editors, *Proceedings of the ACL Workshop on Deep Lexical Acquisition*, pages 28–37. Association for Computational Linguistics, 2005.

DOI: 10.3115/1631850.1631854.

[15] Alina Maria Ciobanu and Anca Dinu. Alternative measures of word relatedness in distributional semantics. In *Joint Symposium on Semantic Processing*, page 80, 2013.

[16] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. NLP (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[17] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA, 2013. ACM. DOI: 10.1145/2506182.2506198.

[18] Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. A tensor-based factorization model of semantic compositionality. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2013)*, pages 1142–1151, Atlanta, GA, US, 2013. Association for Computational Linguistics (ACL).

[19] Cícero Nogueira dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1818–1826, 2014.

[20] Oren Etzioni. Machine reading at web scale. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 2–2. ACM, 2008. DOI: 10.1145/1341531.1341533.

[21] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume One*, IJCAI'11, pages 3–10. AAAI Press, 2011. DOI: 10.5591/978-1-57735-516-8/IJCAI11-012.

[22] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1606–1615, 2015. DOI: 10.3115/v1/N15-1184.

[23] Roxana Girju, Adriana Badulescu, and Dan Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 1–8, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. DOI: 10.3115/1073445.1073456.

[24] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346, 1990. DOI: 10.1016/0167-2789(90)90087-6.

[25] Stevan Harnad. Categorical perception. In L. Nadel, editor, *Encyclopedia of Cognitive Science*, pages 67–4. Nature Publishing Group, 2003.

[26] Stevan Harnad. To cognize is to categorize: Cognition is categorization. *Handbook of categorization in cognitive science*, pages 20–45, 2005.

[27] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954. DOI: 10.1080/00437956.1954.11659520.

[28] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[29] Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 286–295, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[30] Rodolphe Jenatton, Nicolas L. Roux, Antoine Bordes, and Guillaume R Obozinski. A latent factor model for highly multi-relational data. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3167–3175. Curran Associates, Inc., 2012.

[31] Douwe Kiela, Felix Hill, and Stephen Clark. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2044–2048, 2015.

[32] Arne Köhn. What's in an embedding? Analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2067–2073, 2015.

[33] Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4, 2011. DOI: 10.1145/2050104.2050105.

[34] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. DOI: 10.1037/0033-295X.104.2.211.

[35] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[36] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015. DOI: 10.3233/SW-140134.

[37] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):32–38, 1995. DOI: 10.1145/219717.219745.

[38] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2181–2187, 2015. DOI: 10.1016/j.procs.2017.05.045.

[39] H. Liu and P. Singh. ConceptNet –A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226,

October 2004. DOI: 10.1023/B:BTTJ.0000047600.45421.6d.

[40] Max Lungarella, Giorgio Metta, Rolf Pfeifer, and Giulio Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003. DOI: 10.1080/09540090310001655110.

[41] Julian J. McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 785–794, 2015. DOI: 10.1145/2783258.2783381.

[42] John McCarthy. Circumscription - A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1-2):27–39, 1980. DOI: 10.1016/0004-3702(80)90011-9.

[43] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop Papers*, 2013.

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[45] George A. Miller. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. DOI: 10.1145/219717.219748.

[46] Guido Minnen, John A. Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001. DOI: 10.1017/S1351324901002728.

[47] Dipendra K. Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, 2016. DOI: 10.1177/0278364915602060.

[48] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 236–244, 2008. DOI: 10.1039/9781847558633-00236.

[49] Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Tanti Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2302–2310, 2015.

[50] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[51] Raymond J. Mooney. Learning to connect language and perception. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1598–1601, 2008.

[52] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250, 2012. DOI: 10.1016/j.artint.2012.07.001.

[53] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics, 2006. DOI: 10.3115/1220175.1220190.

[54] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 337–346, 2011. DOI: 10.1145/1963405.1963455.

[55] Michaela Regneri, Alexander Koller, and Manfred Pinkal. Learning script knowledge with web experiments. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 979–988, 2010.

[56] Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 109–117, 2010.

[57] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 74–84. The Association for Computational Linguistics, 2013.

[58] Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. Testing apsyn against vector cosine on similarity estimation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016*, pages 229–238, 2016.

[59] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra Kumar Misra, and Hema Swetha Koppula. Robobrain: Large-scale knowledge engine for robots. *CoRR*, abs/1412.0691, 2014.

[60] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 926–934, 2013.

[61] Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, 2012.

[62] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Ma-*

chine Learning Research, 15:1929–1958, 2014.

[63] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706, 2007. DOI: 10.1145/1242572.1242667.

[64] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance multi-label learning for relation extraction. In Jun'ichi Tsujii, James Henderson, and Marius Pasca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 455–465. ACL, 2012.

[65] Leonard Talmy. The fundamental system of spatial schemas in language. *From perception to meaning: Image schemas in cognitive linguistics*, 3, 2005. DOI: 10.1515/9783110197532.3.199.

[66] Moritz Tenorth and Michael Beetz. Knowrob: A knowledge processing infrastructure for cognition-enabled robots. *The International Journal of Robotics Research*, 32(5):566–590, 2013. DOI: 10.1177/0278364913481635.

[67] Moritz Tenorth, Daniel Nyga, and Michael Beetz. Understanding and executing instructions for everyday manipulation tasks from the world wide web. In *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*, pages 1486–1491, 2010. DOI: 10.1109/ROBOT.2010.5509955.

[68] Markus Waibel, Michael Beetz, Raffaello D'Andrea, Rob Janssen, Moritz Tenorth, Javier Civera, Jos Elfring, Dorian Gálvez-López, Kai Häussermann, J.M.M. Montiel, Alexander Perzylo, Björn Schießle, Oliver Zweigle, and René van de Molengraft. RoboEarth - A World Wide Web for Robots. *Robotics & Automation Magazine*, 18(2):69–82, 2011. DOI: 10.1109/MRA.2011.941632.

[69] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 1112–1119, 2014.

[70] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492, 2010. DOI: 10.1109/CVPR.2010.5539970.

[71] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. *ICLR*, 2015.