Semantic Query Processing
Estimating Relational Purity

Kalo, Jan-Christoph; Lofi, Christoph; Maseli, René Pascal; Balke, Wolf-Tilo

Publication date
2017

Document Version
Final published version

Published in
LWDA 2017 Lernen Wissen Daten Analysen 2017

Citation (APA)
Kalo, J.-C., Lofi, C., Maseli, R. P., & Balke, W.-T. (2017). Semantic Query Processing: Estimating Relational Purity. In M. Leyer (Ed.), *LWDA 2017 Lernen Wissen Daten Analysen 2017: Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings* (pp. 113-124). (CEUR Workshop Proceedings; Vol. 1917). CEUR-WS. http://ceur-ws.org/Vol-1917/paper20.pdf

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Semantic Query Processing: Estimating Relational Purity

Jan-Christoph Kalo[1], Christoph Lofi[2], René Pascal Maseli[1] and Wolf-Tilo Balke[1]

[1] Institut für Informationssysteme, TU Braunschweig
r.maseli@tu-bs.de, {kalo, balke}@ifis.cs.tu-bs.de
[2] Web Information Systems Group, Delft University of Technology
c.lofi@tudelft.nl

Abstract. The use of semantic information found in structured knowledge bases has become an integral part of the processing pipeline of modern intelligent information systems. However, such semantic information is frequently insufficient to capture the rich semantics demanded by the applications, and thus corpus-based methods employing natural language processing techniques are often used conjointly to provide additional information. However, the semantic expressiveness and interaction of these data sources with respect to query processing result quality is often not clear. Therefore, in this paper, we introduce the notion of *relational purity* which represents how well the explicitly modelled relationships between two entities in a structured knowledge base capture the implicit (and usually more diverse) semantics found in corpus-based word embeddings. The purity score gives valuable insights into the completeness of a knowledge base, but also into the expected quality of complex semantic queries relying on reasoning over relationships, as for example analogy queries.

Keywords: Semantics of Relationships, LOD, Structured Knowledge Repositories, Word Embeddings

1 Introduction

To provide an intuitive and efficient user experience, future information systems need to offer powerful query capabilities with an awareness of the semantics of the query, as for example question answering systems [1] or intelligent digital assistants like MS Cortana, Google Now, or Apple Siri. Here, entities referred to in queries and relationships between those entities play a central role in the query processing process. *Structured Knowledge Repositories* like the Google Knowledge Vault [2] or WordNet [3] and *Linked Open Data* sources like DBpedia [4], Yago [5], usually serve as a premier source of such semantic information. However, due to their nature as structured knowledge bases, they are not always sufficient to implement some concepts required for intuitive queries: for example, information on human perception (like similarity or relatedness) or information on less clear attributes or fuzzy relationships are often omitted. As an example, consider the entities *Brad Pitt* and *Angelina Jolie*. In addition to their only (and outdated) DBpedia relationship *spouse*, their semantic relationship of

In: M. Leyer (Ed.): Proceedings of the LWDA 2017 Workshops: KDML, FGWM, IR, and FGDB.
Rostock, Germany, 11.-13. September 2017, published at http://ceur-ws.org

course is much more complex. There exists a plethora of additional relationships between them, which is quite hard to model: as for example, they co-acted in the same movies together, had many joint public appearances, and publicly split up again.

Here, *word embeddings* (such as recent skip-n-gram, neural embeddings [6, 7]) have been shown to provide an interesting additional source of semantic information on entities and their relationships: word embeddings learn a vector representation of words used in a large natural language corpus by exploiting the distributional hypothesis [8], thus promising to encode semantics based on the actual human perception and everyday use, implicitly provided by the language structure of the natural text corpus used for training (like news articles or encyclopedias). For example, it has been shown that the similarity of the resulting word vectors closely correlates with the perceived attributional similarity of their respective real-world entities [9], which allows for similarity queries but also for mapping between diverging user and knowledge base vocabulary for query processing in question answering [10] (i.e., when querying for "husband", but the knowledge base only has information on "spouses", similarity can help to suggest using spouse instead of husband).

An interesting use case of powerful semantic queries which were (re-)popularized by word embeddings are analogy queries. Using analogies in natural speech allows communicating dense information easily and naturally by implying that the "essence" of two concepts is similar or at least perceived similarly. Thus, analogies can be used to map factual and behavioral properties from one (usually better-known concept, the source) to another (usually less well known, the target) concept by exploiting both attributional and relational similarity. This is particularly effective for natural querying and explaining when only vague domain knowledge is available (e.g., "Okinawa is to Japan as is Hawaii to the US"). It has been argued [6] that such semantics can be expressed by simple vector arithmetics within the word embeddings space, however, it has also been shown that the performance of this type of analogy processing varies greatly with the type of relationships involved [11].

In this paper, based on contemporary computational analogy processing theory, we investigate the semantic relationship between vector arithmetics of word embeddings, the relationships found in structured knowledge repositories, and analogy semantics. The resulting contributions can be summarized as follows:

- We introduce the concept of *purity scores* for relationships in knowledge bases by investigating the vectors associated with each relationship in a word embeddings space. As word embeddings are based on rich language semantics of a large text corpus, we assume that they contain richer (or at least different) semantic information than structured knowledge bases, but only in implicit form. The purity score of a relationship represents the degree to which the knowledge base covers the implicit semantics of embeddings.
- We provide an extensive overview and examples of different relationships, their associated vectors, as well as their related source texts which have been involved in creating those vectors to clarify the concept of purity scores
- We show that the popular analogy reasoning technique using word embedding vector arithmetics work well for pure relationships, while it does not work well for impure ones.

2 Related Concepts

In the following section, we revisit and summarize some of the core concepts underpinning our findings. This especially covers the general semantics of 4-term analogies, common word embeddings, and the offset method for analogical reasoning using vector arithmetics (as we already discussed in [12]).

Analogies and Relational Similarity. The semantics of analogies have been researched in depth in the fields of philosophy, linguistics, logics, and in cognitive sciences, such as [13–15]. However, those models are rather complex and hard to grasp computationally, and thus most recent works on computational analogy processing rely on the simple 4-term analogy model, which is given by two sets of word pairs (the so-called analogons), with one pair being the source and one pair being the target. A 4-term analogy holds true if there is a high degree of relational similarity between those two pairs. This is denoted by $[a_1, a_2] :: [b_1, b_2]$, where one relationship between a_1 and a_2 is similar to a relationship between b_1 and b_2, as for example in [US Dollar, USA]::[Euro, Germany]. This model has several limitations, as is discussed in [16]: the semantics of "a high degree of relational similarity" from an ontological point of view is unclear as there can be plethora of relationships between the concepts of an analogon, but only some of them are of relevance for valid analogy semantics.

Therefore, we rely on an improved interpretation of the 4-term analogy model [16], and assume that there can be multiple relationships between the concepts of an analogon, some of them being relevant for the semantics of an analogy (the *defining* relationships), and some of them not. An analogy holds true if the sets of defining relationships of both analogons show a high degree of relational similarity. For illustrating the difference and importance of this change in semantics, consider the analogy statement: [Tokyo, Japan]::[Braunschweig, Germany]. Tokyo is a city in Japan, and Braunschweig is a city in Germany, therefore both analogons contain the same "city is located in country" relationship, and this could be considered a valid analogy with respect to the simple 4-term analogy model. Still, this is a poor analogy statement from a human perspective because Braunschweig is not like Tokyo at all (therefore, this statement does neither describe the essence of Tokyo nor the essence of Braunschweig particularly well): the defining traits (relationships) of Tokyo in Japan should at least cover that Tokyo is the single largest city in Japan, and its capital. There are many other cities, which are also located in Japan, but only Tokyo has these two defining traits. Braunschweig, however, is just a smaller city in Germany, which might stand out for either its technical university or its historic city center (therefore, the defining relationships of both word pairs are not very similar). The closest match to a city like Tokyo in Germany should therefore be Berlin, which is also the largest city and the capital city.

Understanding which relationships define the essence of an analogon as perceived by humans is a very challenging problem, but this understanding is crucial for judging the usefulness and value of an analogy statement. Furthermore, the degree in which relationships are defining an analogon may vary with different contexts (e.g., the role of Berlin in Germany in a political discussion vs. the role of Berlin in Germany in a discussion about nightlife).

Word Embeddings, Relational Similarity, and Analogy Processing. Word embeddings represent each word in a predefined vocabulary with a real-valued vector, i.e. words are embedded in a vector space (usually with 50-600 dimensions). Most word embeddings will directly or indirectly rely on the distributional hypothesis [8] (i.e. words frequently appearing in similar linguistic contexts will also have similar real-world semantics), and are thus particularly well-suited to measure semantic similarity and relatedness between words (which is one of the foundation of the 4-term analogy definition), e.g., see [9]. In recent years, especially word embeddings relying on neural networks have become popular, with the skip-gram negative sampling approach (SGNS) [6, 7] being one of the best known examples .

The straight-forward application of word embeddings is computing similarity between two given words [9] by measuring the cosine similarity. However, many (but not all) word embeddings show some very interesting and surprising additional property: it seems that not only the cosine distance between vectors represents a measure for similarity and relatedness of the embedded words, but that also the difference vectors between word pairs implicitly represent the relationships between two entities, and thus carry analogy semantics [17]. For example, the difference between the vector for "man" and "king" seems to represent the concept/relationship of being a ruler, and the closest word vector to "woman" plus the "ruler" concept vector will be "queen" (see Fig. **1**; this method is also sometimes called the *offset method*).

To a certain extent, these semantics can be attributed to the distributional hypothesis: in natural speech, concepts carrying similar semantics will frequently co-occur in similar context. Therefore, the difference vector should implicitly encode the defining relationships between two concepts as discussed in the previous section (i.e.: Toyko/Japan and Berlin/Germany will likely occur in similar contexts in natural speech, while Braunschweig/Germany will likely appear in different context and will thus have a different difference vector). The ability of word embeddings to perform this analogical reasoning process has been evaluated using several standardized test sets (see next section), but is still not well understood and can fail quite often, which is related to our introduced *purity score*.

In a more formal fashion, a word embedding can be used to solve *analogy completion queries* as follows [6]: Given the query $[a_1, a_2] :: [b_1, ?]$, the word embedding provides the respective word vectors $\overrightarrow{a_1}$, $\overrightarrow{a_2}$, and $\overrightarrow{b_1}$. Then, the vector $\overrightarrow{b_2}$ representing the query's solution can be determined by finding the word vector in the trained vector

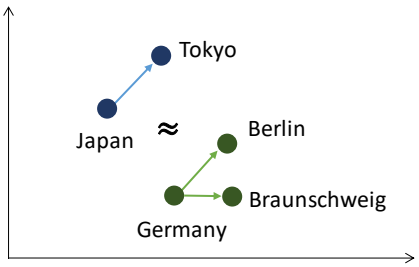

Fig. 1. Schematic representation from our GloVe Word Embedding Vectors for the DBpedia relationship *country*.

$$\overrightarrow{Tokyo} - \overrightarrow{Japan} + \overrightarrow{Germany} \approx \overrightarrow{Berlin}$$

space V which is closest to $\overrightarrow{a_2} - \overrightarrow{a_1} + \overrightarrow{b_1}$ with respect to the cosine vector distance, i.e. $\overrightarrow{b_2} = \arg \max_{\vec{x} \in V, \vec{x} \neq \overrightarrow{a_2}, \vec{x} \neq \overrightarrow{b_1}} \left(\overrightarrow{a_2} - \overrightarrow{a_1} + \overrightarrow{b_1}\right)^T \vec{x}$.

Relational Benchmarking. The Mikolov Benchmark set [18] is one of the most popular benchmark sets for testing the analogy reasoning capabilities of word embeddings, covering 19,558 4-term analogies. However, only 14 distinct relationships are covered, and most of them (9) focus on grammatical properties the relationship "is plural for a noun", e.g., [mouse,mice]::[dollar,dollars] or "is superlative". Five relationships are of a semantic nature (i.e. "is capital city for country" [Athens,Greece]::[Oslo,Norway], "is currency of country", "city in state", "male-female version" (including the often cited [king,queen]::[man,woman]). The test set is generated by collecting pairs of entities which are members of the selected relationship either manually or from Wikipedia and DBpedia, and then combining these pairs into 4-term analogy tuples. For example, for the "city in state" relationship, 68 word pairs like [Dallas,Texas] or [Miami,Florida] are collected, and then combined by a cross product. The related Wordrep dataset [19] extends the Mikolov set by adding more challenges, and expanding to 25 different relationships.

For the Mikolov dataset, the authors showed that skip n-gram word embeddings [6] using the offset method could solve analogy completion queries (i.e. [Athens,Greece]::[Oslo,?]) with an accuracy of 53.3% overall, 50% for semantic relationships (like 'capital of'), and 55.9% for syntactic ones (like 'plural of'). No deeper analysis of the relationships for which this technique performs well was provided. However, this was analyzed in more detailed in [11] using subjective feedback on relational similarity from human users. Here, the authors identified as the core problem which hinders the offset methods for analogy reasoning the presence of multiple relationships between two entities which influence human perception, as some relationships are perceived more dominant than others, e.g., quite often, the relationship intended in the Mikolov analogy challenge (e.g., "city in state") was not perceived as dominant as some other relationship perceived by human subjects (e.g., "home of best football team in state"). While this argument is formulated slightly differently, those experimental results strongly support the hypothesis of defining relationships [16] discussed in the previous sections. Also, different perception of relationships introduces problems with symmetry or transitivity of relational similarity not holding from a user's perspective [11]. A similar result is also supported by experiments in [20].

Based on the intuition obtained in those experimental results, in the following, we define the concept of *relational purity* approximating in how far a relationship given in an analogy challenge is indeed perceived as the relevant or defining relationship, which gives us insights both into the its suitability for analogy query processing, but can also serve as an indirect and implicit measure for the semantic completeness of a knowledge base with respect to that relationship.

3 Purity Score for Relationships

In the following section, we further explain our idea of relationship purity and provide a formal definition of the concept. We generalize the idea to relationships between entities, and support our findings with an analysis of DBpedia relationships.

Motivation. As we motivated, the semantic relationships modelled in triple format (consisting of subject, predicate, object triples) in state-of-the-art knowledge bases are often a stark simplification of the relationships between the respective entities as perceived by humans. As a result, many relationships are left out (e.g., Angelina Jolie and Brad Pitt having a public fight about their children), or several related but still perceptually different relationships are generalized and grouped into a single relationship. In Figure 2, we have visualized a 2-dimensional scaling of a GloVe embedding trained on Wikipedia for the "spouse" relationship (i.e., the difference vectors between two spouses). The more parallel those relationship vectors are, the more similar their representation is in the embedding space, and therefore, the more similar their captured semantics should be (as the texts in which those relationships are discussed share similar contexts). In the case of the spouse relationship, we observe that the relationship vectors of different couples are quite diverse: However, we can observe that the vectors of the two actor couples (Jolie/Pitt and Heard/Depp), are more similar, since they are linked by more than one single relationship.

If, in DBpedia, we look at the entities *Tolkien* and *The Hobbit*, we observe that they are connected by only a single relationship instance: Tolkien is the author of The Hobbit. In the text found in Wikipedia, a whole paragraph is used to describe the relationship between the respective entities: how he initially wrote The Hobbit for his children, how he never planned to publish it, how his friends liked it and pressured him towards publication, etc. Similarly, the same "is author of" relationship is used to link *Max Frisch* to his novel *Homo Faber*, whereas the Wikipedia text offers a much deeper insight into their relationship than the structured knowledge base does (see Fig. 2 for a visualization).

However, as defined by the DBpedia ontology, the recommended use of the "is author of" relationship is much more general and it can connect any type of author with their work of any kind (this can be novels, scientific papers, screenplays, music, paintings and even software programs). Thus, it generalizes the semantics of a larger amount of diverse *subrelationships* to one single relationship, not capturing their perceived semantic differences anymore (i.e., Pablo Picasso authoring the painting "Les Demoiselles d'Avignon" is perceived quite different than Max Frisch authoring "Homo Faber" by most). However, we believe that this loss of semantics can be modelled and captured by analyzing rich textual corpora. Here, for the "is author of" relationship, we claim that the rich diversity of different types of author relationship leads to a low *purity score*, while other relationships like "is currency of country" have high purity scores (while there might be diverse relationships between countries and their currencies, this diversity is comparably low). This *purity*, i.e., diversity of usage of a relationship can be observed in word embeddings.

As a further example relationship in Fig. 2, consider the "is in country" relationship for cities: the relationship between Braunschweig and Germany for example (i.e., there

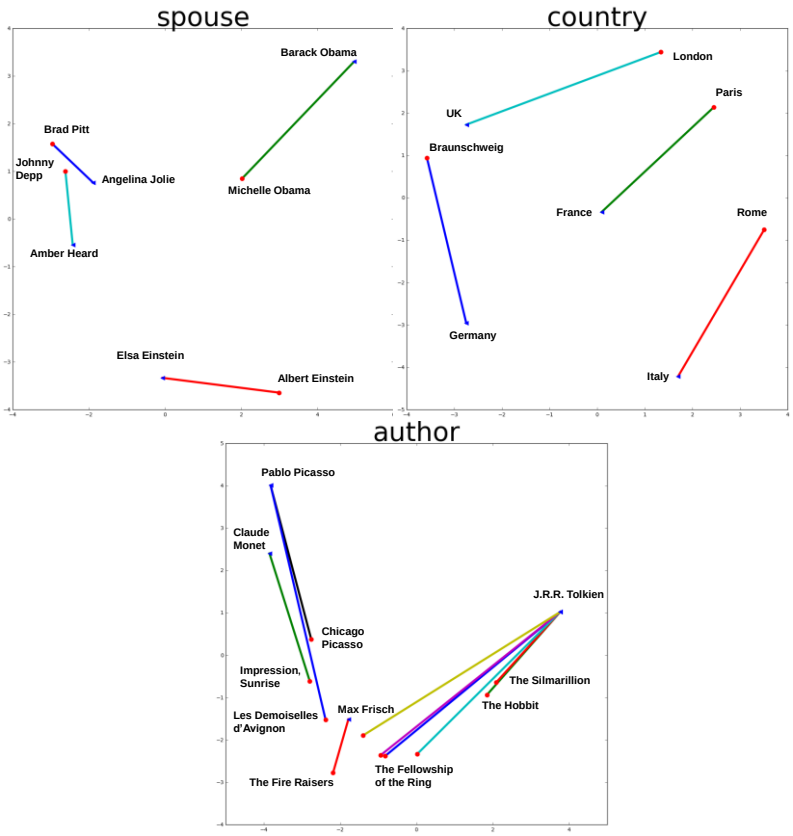

Fig. 2. Three DBpedia relationships in the embedding space visualized in 2D using Principal Component Analysis.

is nothing particularly unique about Braunschweig), is quite different from the relationship from Paris to France. Paris is not only a city located within France, but also the country's largest city, the location of the French government and its capital. The relationship vector of Rome to Italy and London to UK look very similar.

Please note that all those vectors have 300-dimensions, and that thus such simple visual analytics are quite crude due to the loss of semantics when mapped to the 2D-plane using Principal Component Analysis. Hence, we introduce a formal definition of relational purity in the next section.

Computing the Purity Score. Word embeddings were shown to provide linear substructures that can represent implicit relational similarity of all relationships between two entities as having similar (i.e., having a high cosine similarity) difference vectors between their word vectors. This characteristic is mainly used for analogical query processing using the offset method. However, we can adopt a similar notion to compute the *purity score* of relationships. Given a set subject entities S and object entities O, which are connected by a relationships R in a structured knowledge base, we compute

the purity score of the relationship $R \subseteq S \times O$ as the standard deviation from the average difference vector of the entities. Given a triple (s, r, o), its relationship vector in a word embedding is defined by the respective entities difference vector: $\vec{r} = \vec{s} - \vec{o}$. We define the *average vector* for relationship R, given the set of relation vectors $\vec{R}$ as $\overrightarrow{a_R} = \frac{\sum_{\vec{r} \in \vec{R}} \vec{r}}{|\vec{R}|}$. Now, we define the purity score of a relationship as the standard deviation of the cosine distance from every relationship instance vector to the average vector $\overrightarrow{a_R}$. The cosine distance between two vectors is defined as $\cos(u, v) = 1 - \frac{\vec{u} \cdot \vec{v}}{\|u\|_2 \cdot \|v\|_2}$. Note that negative similarity leads to distances between 1 and 2 in case the vectors are directed in opposite directions. The purity of a relationship R is now defined as:

$$pur(R) = 1 - \frac{\sum_{\vec{r} \in \vec{R}} cos(\overrightarrow{a_R}, \vec{r})^2}{|\vec{R}|}$$

A high variance in the directions of the relationship vectors (represented by cosine distances to the average vector), results in impure relational information in the embedding, i.e. low purity scores. Similarly, a low variance in the directions (so parallel relationship vectors) lead to a high purity score.

4 Evaluation

In this section, we introduce our experimental setup, and specifically focus on how we represented DBpedia entities and relationships in a corpus-based word embedding. Afterwards, we compute the purity scores for DBpedia relationships and visualize and discuss some examples to obtain a better intuition of the results.

Experimental Setup. We extracted relationships between entities from the largest Linked Open Data (LOD) data store DBpedia [4], a knowledge graph that is built by extracting knowledge from Wikipedia Infoboxes. As result, our dataset covers 18 million unique relationship instances between entities from around 1,200 relationships as defined by the DBpedia ontology, ranging from capital city relationships, over causal relationships, to biological relationships between living organisms.

As a training corpus for our word embedding, we downloaded a dump of the English version of Wikipedia from 01/2017. For linking DBpedia entities and relationships to the embedding, as a first step, we performed named entity recognition and disambiguation on the text corpus using DBpedia Spotlight [21] and replace the recognized entities by their respective DBpedia URI. For creating the word embedding, we use GloVe [7] and train multiple models with varied window size as described in the original GloVe paper between 8 and 15. Similar to their results [7], we found out that semantic relationships are better represented in the embedding when we use the larger window size. Furthermore, we ensured that the sliding window does not reach over different articles and sentence boundaries, preventing words appearing in wrong contexts. We varied the number of embedding dimensions between 50, 100 and 300, finding that DBpedia relationships are represented best by choosing 300 dimensions. Thus, the fol-

lowing results all are based on a GloVe embedding with window size 15, 300 dimensions and 100 training iterations. The minimum word frequency is set to 8. Our corpus and embeddings are available on request.

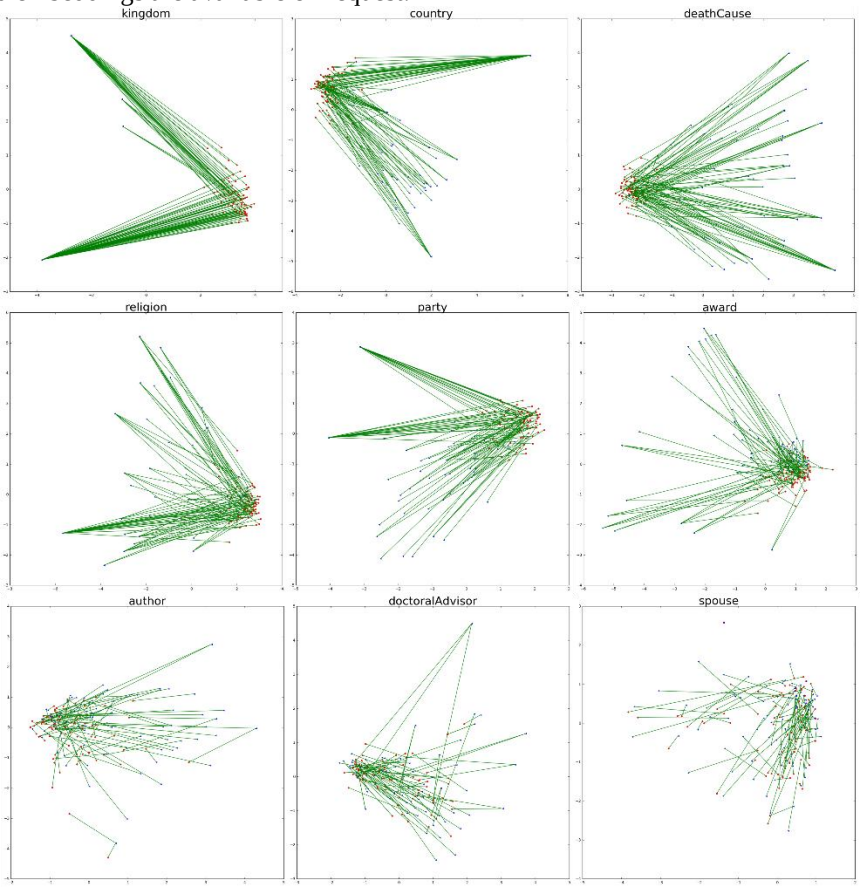

Fig. 3. Two-dimensional visualization of randomly sampled difference vectors for 9 DBpedia ontology relationships. The projection was computed using principal component analysis. The relationships are ordered by purity from top-left to bottom-right. Further information on the relationships can be seen in Table 1.

Result Visualization. For visualizing relationship vectors and their purity, we selected 9 DBpedia relationships from very pure to extremely impure relationships from a 300-dimensional GloVe embedding. We used principal component analysis to scale the results to two dimensions. The results are visualized in Figure 3. Subject entities are visualized by a red dot, object entities by a blue triangle and the difference vector, representing the relationship between the entities, is visualized by a green line. The relational similarity between the relationships given by cosine distance of the difference vectors. Hence, parallel vectors have cosine similarity 0, whereas orthogonal vectors have similarity 1. (Due to the two-dimensional scaling of the original vectors, the cosine

similarity is not perfectly represented in the visualization.) Pure relationships (top left) have highly parallel relationship vectors, whereas impure relationships (bottom right) are very diverse. Since impure relationships have nearly no parallel relationship vectors, the offset method does not lead to meaningful results. Hence, most analogy queries based on this method return incorrect results.

Table 1. DBpedia relationships, their purity score and the domain and range of the respective relationships.

Relationship	***Purity***	***Range, Domain***	***Relationship***	***Purity***	***Domain, Range***
kingdom	0.94	Species, Taxonomic Rank	*currency*	0.56	Country, Currency
mediaType	0.93	Book, Media Type	*musicComposer*	0.53	Composition, Composer
domain	0.92	Species, Taxonomic Rank	*musicalBand*	0.49	Song, Music Band
gender	0.90	Person, Gender	*director*	0.47	Movie, Director
phylum	0.89	Species, Taxonomic Rank	*award*	0.46	Person, Award
timeZone	0.82	City, Timezone	*coach*	0.46	Sports Team, Coach
country	0.81	City, Country	*capital*	0.43	Country/State, Capital
sport	0.80	Organization, Sport	*writer*	0.41	Book/Screenplay, Writer
nationality	0.80	Person, Nationality	*editor*	0.38	Book/Movie, Editor
profession	0.79	Person, Job	*author*	0.36	Book/Painting/Software/Movie, Author
deathCause	0.79	Person, Cause of Death			
campus	0.77	University, Campus			
occupation	0.76	Person, Job	*album*	0.34	Song, Music Album
religion	0.75	Person, Religion	*doctoralAdvisor*	0.24	Person, Person
instrument	0.74	Person, Instrument	*child*	0.23	Person, Person
hometown	0.67	Music Artist, City	*parent*	0.09	Person, Person
mayor	0.67	City, Person/Political Party	*partner*	0.07	Person, Person
party	0.66	Person, Political Party	*relative*	0.04	Person, Person
university	0.58	Person, University	*spouse*	0.01	Person, Person

Purity of DBpedia Relationships. We have evaluated the purity of all DBpedia ontology relationships for which we could find at least two relationship instances in our embedding. This resulted in more than 400 different relationship embeddings with different purity scores. In Table 1, we show an excerpt covering the complete purity spectrum. Particularly pure are relationships with only a few different subjects or objects, as for example, the biological kingdom relationship from DBpedia which connects species to one of six different biological kingdoms. The country relationship linking cities to their countries or states also has a high purity score of 0.81: most of the relationships instances are parallel, having some exceptions as shown in Figure 3. The author relationship, as already discussed in Section 3, has a purity of 0.36. Since its domain comprises entities of very different type, the resulting relationship vectors show only few similarities. However, we can see several different clusters of similar (parallel) relationships, indicating that the relationship is impure (i.e., each cluster represents a subrelationship which is not modelled directly in the knowledge base).

The spouse relationship connecting two married persons is the most diverse (and thus impurest) relationship in our dataset, having a purity score of only 0.01. This diversity has several reasons: On the one hand, this relationship is symmetric (which is not covered by the default similarity measurement using cosine distance), therefore the vectors for man and his wife is directed in the opposite direction to the vector that con-

nects a woman to her husband. Furthermore, the persons being married and their relationships to each other are quite different from couple to couple (see our introductory examples) which is well represented in text but not in a knowledge base.

5 Summary and Future Work

In this paper, we investigated the semantic interplay between explicit relationships modelled in structured knowledge bases like DBpedia, and their representation in corpus-based word embeddings. While word embeddings do not explicitly represent individual relationship instances between two entities, the implicit representation of all relationship instances as a difference vectors between two word/entity vectors can potentially cover a much wider range of (perceptual) semantics based on the corpus used for training. This notion is usually exploited for embedding-based relational similarity and analogy queries. However, it can also be used to shed some light on the nature of a specific relationship and how well it is represented in a knowledge base. To this end, we introduced the concept of *relational purity*, which implicitly represents how uniform the usage of a give relationship is in natural text. This results on several interesting observations: some relationships (like "is spouse of") carry much richer semantics in their textual representations (e.g., while DBpedia just contains that Brad Pitt and Angelina Jolie are/used to be married, texts talking about both indicate a large number of additional potential and currently not covered relationships) indicated by a low purity score, while hierarchical relationships from the Linné taxonomy for animal and plant life are very pure – e.g., indicating that there is exactly one specific semantic relationship between two entities in the taxonomy (for example, squirrels are rodents – there are no significant other relationships mentioned in natural text connecting "squirrel" with "rodent").

For the future, we plan to exploit our insights obtained in this work for improving complex semantic query processing (by e.g., being able to give an assessment of the potential reliably of an answer based on reasoning over relationships), but also for designing processes for uncovering potential semantic gaps in knowledge bases and mining for missing information in a targeted fashion.

References

1. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine. 31, 59–79 (2010).
2. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining (SIGKDD). pp. 601–610. ACM Press, New York, New York, USA (2014).
3. Miller, G.A., A., G.: WordNet: a lexical database for English. Communications of the ACM. 38, 39–41 (1995).
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus

for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference (ISWC). pp. 722–735. Springer Berlin Heidelberg (2007).
5. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web - WWW '07. p. 697. ACM Press, New York, New York, USA (2007).
6. Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language (NAACL-HLT). pp. 746–751. , Atlanta, USA (2013).
7. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014).
8. Harris, Z.: Distributional Structure. Word. 10, 146–162 (1954).
9. Lofi, C.: Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approaches. Database Society of Japan (DBSJ) Journal. 14, 1–9 (2016).
10. Freitas, A., Faria, F.F. de, Seán O'Riain, Curry, E.: Answering natural language queries over linked data graphs: a distributional semantics approach. In: Proceedings of the International Conference on Conference on Research and Development in Information Retrieval (SIGIR). , Dublin, Ireland (2013).
11. Chen, D., Peterson, J.C., Griffiths, T.L.: Evaluating vector-space models of analogy. In: Proceedings of the Conference of the Cognitive Science Society. , London, UK (2017).
12. Lofi, C., Ahamed, A., Kulkarni, P., Thakkar, R.: Benchmarking Semantic Capabilities of Analogy Querying Algorithms. In: Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA). pp. 463–478. , Dallas, TX, USA (2016).
13. Dedre Gentner, Keith J. Holyoak, Boicho N. Kokinov eds: The analogical mind: perspectives from cognitive science. MIT Press (2001).
14. Shelley, C.: Multiple Analogies In Science And Philosophy. John Benjamins Pub. (2003).
15. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive science. 7, 155–170 (1983).
16. Lofi, C., Nieke, C.: Modeling Analogies for Human-Centered Information Systems. In: Jatowt, A., Lim, E.-P., Ding, Y., Miura, A., Tezuka, T., Dias, G., Tanaka, K., Flanagin, A., and Dai, B.T. (eds.) Social Informatics (SocInfo). pp. 1–15. Springer International Publishing, Cham (2013).
17. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research. 2493–2537 (2011).
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR. abs/1301.3, (2013).
19. Gao, B., Bian, J., Liu, T.-Y.: WordRep: A Benchmark for Research on Learning Word Representations. In: ICML Workshop on Knowledge-Powered Deep Learning for Text Mining. , Beijing, China (2014).
20. Linzen, T.: Issues in evaluating semantic spaces using word analogies. In: Workshop on Evaluating Vector Space Representations for NLP. , Berlin, Germany (2016).
21. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11. pp. 1–8. ACM Press, New York, New York, USA (2011).