UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

WeST People and Knowledge Networks
Institute for Web Science and Technologies

Enrichment of ontological taxonomies using a neural network approach

Bachelorarbeit

zur Erlangung des Grades einer Bachelor of Science (B.Sc.)
im Studiengang Informatik

vorgelegt von
Alex Baier

Erstgutachter: Prof. Dr. Steffen Staab
Institute for Web Science and Technologies

Zweitgutachter: Max Mustermann
Institute for Web Science and Technologies

Koblenz, im Januar 2017

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

	Ja	Nein
Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden.	☐	☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	☐	☐
Der Text dieser Arbeit ist unter einer Creative Commons Lizenz verfügbar.	☐	☐
Der Quellcode ist unter einer Creative Commons Lizenz verfügbar.	☐	☐
Die erhobenen Daten sind unter einer Creative Commons Lizenz verfügbar.	☐	☐

. . . .

(Ort, Datum) (Unterschrift)

Contents

1 Introduction **1**

2 Foundations **1**
2.1 Wikidata 1
2.2 Taxonomy and problem definition 1
2.3 Similarity 4
2.4 Similarity-based classification 4
2.5 Text processing 5

3 Analysis of the Wikidata taxonomy **5**

4 Ontology learning **5**

5 Neural networks **6**
5.1 Recursive neural networks for graph representation 6
5.2 Deep neural networks for graph representation 6
5.3 Continuous Bag-of-Words 6
5.4 Skip-gram with negative sampling 6
5.5 Comparison 6

6 Algorithm **6**
6.1 Baseline 6
6.2 Supplementing with other resources 6

7 Evaluation **6**
7.1 Method 6
7.2 Generation of gold standard 7
7.3 Results 7

List of Figures

1 Example for k-nearest neighbors for 3 classes with k=4 and k=10. . . 5

1 Introduction

Motivation. Related work. Solution. Evaluation.

2 Foundations

2.1 Wikidata

TODO: Define entity in Wikidata, how are classes identified, etc. Galárraga [9] Wikidata is a open, collaborative and user-driven knowledge base. Its main purpose is to serve as a structural knowledge store for other Wikimedia projects like Wikipedia.

2.2 Taxonomy and problem definition

TODO: maybe separate taxonomy and problem definition in to two sections?

- Ontology
  Cimiano [4]
  Galárraga [9]
- Taxonomy
  Cimiano [4]
  Galárraga [9]
- Connected taxonomy (maybe: consistent taxonomy)
- Root class
- Unlinked class
- Problem statement

TODO: General notion of ontology and taxonomy
Cimiano [4] defines ontology, which includes the taxonomy, as follows:
"

Definition (Ontology). *An* ontology *is a structure*

$$\mathcal{O} := (C, \leq_C, R, \sigma_R, \leq_R, \mathcal{A}, \sigma_\mathcal{A}, \mathcal{T})$$

consisting of

- *four disjoint sets C, R, $\mathcal{A}$, and $\mathcal{T}$ whose elements are called* concept identifiers, relation identifiers, attribute identifiers *and* data types, *respectively,*

- *a semi-upper latice $\leq_C$ on C with top element $root_C$, called* concept hierarchy *or* taxonomy,
- *a function $\sigma_R : R \to C^+$ called* relation signature,
- *a partial order $\leq_R$ on R, called* relation hierarchy, *where $r_1 \leq_R r_2$ implies $|\sigma_R(r_1)| = |\sigma_R(r_2)|$ and $\pi_i(\sigma_R(r_1)) \leq_C \pi_i(\sigma_R(r_2))$, for each $1 \leq i \leq |\sigma_R(r_1)|$, and*
- *a function $\sigma_{\mathcal{A}} : \mathcal{A} \to C \times \mathcal{T}$, called* attribute signature,
- *a set $\mathcal{T}$ of datatypes such as strings, integers, etc.*

Hereby, $\pi_i(t)$ is the i-th component of tuple t. [...] Further, a semi-upper lattice $\leq$ fulfills the following conditions:

$\forall x x \leq x$ (reflexive)
$\forall x \forall y (x \leq y \wedge y \leq x \implies x = y)$ (anti-symmetric)
$\forall x \forall y \forall z (x \leq y \wedge y \leq z \implies x \leq z)$ (transitive)
$\forall x x \leq top$ (top element)
$\forall x \forall y \exists z (z \geq \wedge z \geq y \wedge \forall w (w \geq x \wedge w \geq y \implies w \geq z))$ (supremum)

So every two elements have a unique most specific supremum. ”

A taxonomy can be modeled as a semi-upper lattice. This induces two important assumptions about the structure and to some degree completeness of the observed taxonomies. First, there is only one *root class*, top element of the lattice, of which every other class is (transitively) a subclass. Second, because of the supremum property, the taxonomy is fully connected, which means each class, but the root class, has a superclass. Wikidata's taxonomy does therefore not fulfill the definition by Cimiano [4], as it is not fully connected.
In the following, new definitions will presented, which attempt to model an incomplete taxonomy based on the already presented data model and structure of Wikidata. First, basic concepts of graphs will be introduced.

Definition 1 (Directed graph)**.** *A* directed graph *G is an ordered pair $G = (V, E)$, where V is a set of vertices, and $E = \{(v_1, v_2) \mid v_1, v_2 \in V\}$ is a set of ordered pairs called directed edges, connecting the the vertices.*

Definition 2 (Predecessor)**.** *Let $G = (V, E)$ be a directed graph. $v_1 \in V$ is a* predecessor *of $v_2 \in V$, if there exists an edge so that $(v_1, v_2) \in E$. Let $v \in V$ be a vertice of G, then $pred_G(v) = \{w \mid (w, v) \in E\}$ is the set of predecessors of v.*

Definition 3 (Successor)**.** *$v_1 \in V$ is a* successor *of $v_2 \in V$, if there exists an edge so that $(v_2, v_1) \in E$. Let $v \in V$ be a vertice of G, then $succ_G(v) = \{w \mid (v, w) \in E\}$ is the set of successors of v.*

Definition 4 (Walk). *Let $G = (V, E)$ be a directed graph. A* walk *W of length $n \in \mathbb{N}$ is a sequence of vertices $W = (v_1, \ldots, v_n)$ with $v_1, \ldots, v_n \in V$, so that $(v_i, v_{i+1}) \in E \; \forall i = 1, \ldots, n-1$.*

Definition 5 (Cycle). *A walk $W = (v_1, \ldots, v_n)$ of length n is called a* cycle, *if $v_1 = v_n$.*

Definition 6 (Directed acyclic graph). *A directed graph G is called* directed acyclic graph, *if there are no cycles in G.*

In Wikidata, a class can have multiple superclasses, therefore a tree structure is not sufficient to model the taxonomy. However a directed acyclic graph, can model the taxonomy. The acyclic constraint is necessary to ensure that no class is transitively a subclass of itself.

Definition 7 (Taxonomy). *A* taxonomy *$T = (C, S)$ is a directed acyclic graph, where C is a set of* class identifiers, *and S is the set of edges, which describe the* subclass-of relation *between two classes. such that c_1 is the subclass of c_2, if $(c_1, c_2) \in S$.*

Definition 8 (Subclass-of relation). *The transitive binary relation $\lhd_T$ on the taxonomy $T = (C, S)$ represents the subclass relationship of two classes in T. Given $c_1, c_2 \in C$, $c_1 \lhd_T c_2$, if there is a walk $W = (c_1, \ldots, c_2)$ with length $n \geq 1$, which connects c_1 and c_2. $\lhd_T$ is transitive, $\forall c_1, c_2, c_3 \in C : c_1 \lhd_T c_2 \land c_2 \lhd_T c_3 \implies c_1 \lhd_T c_3$.*

If the taxonomy defined by Cimiano [4] is mapped on this graph-based taxonomy model, the following assumption is true, for $T = (C, S)$:

$$|\{c \in C \mid \neg \exists s \in C : c \lhd_T s\}| = 1$$

Only one class in this taxonomy has no superclasses. This class is called *root class*. However in the case of Wikidata, this assumption does not hold true. The following state is the case:

$$|\{c \in C \mid \neg \exists s \in C : c \lhd_T s\}| > 1$$

There are classes other than the root class, which also have no superclasses. These classes will be called *unlinked classes*.

Definition 9 (Root class). *Given a taxonomy $T = (C, S)$, the* root class *$root_T$ is a specific, predefined class with no superclasses in T. For $root_T$, $|succ_T(root_T)| = 0$ applies.*

Definition 10 (Unlinked class). *Given a taxonomy $T = (C, S)$ with a root class $root_T$, a class $u \in C$ is called* unlinked class, *if $u \neq root_T \land |succ_T(u)| = 0$.*

In Wikidata the root class is *entity (Q35120)*, described as something that exists. **TODO: can i cite the Wikidata discussion, where it is said that entity is considered the root node?** The task of this thesis is the classification of unlinked classes in Wikidata. In other words a function is needed, which given an unlinked class u of a taxonomy $T = (C, S)$ with a root class $root_T$, find an appropriate superclass for T. Doan et al. [7] suggests that for the task of placing a class into an appropriate position in T, either finding the most similar class, most specific superclass, or most general subclasses of u, are sensible approaches. This induces that the appropriate superclass for an unlinked class u is either the most similar class $c \in T$, or one of the superclasses of $succ_T(c)$. Therefore we can define the problem, as follows:

Definition 11 (Problem definition)**.** *Given a taxonomy $T = (C, S)$ with root class $root_T$ and a similarity function sim over T, find a function f, which, given an unlinked class $u \in C$, returns a class $s = f(u)$, fulfilling the following criteria:* ***TODO: define as the parents of the most similar class? for example german would be similar to english, therefore the superclass for german should be language and not english***

$$\neg(s \lhd_T u) \text{ no child} \quad (1)$$

$$s = \max_{c \in C}(sim(u, s)) \text{ most similar class} \quad (2)$$

2.3 Similarity

- semantic similarity e.g. distributional similarity
  Lin [15]
  Rodríguez and Egenhofer [19]
- geometrical similarity e.g. distance based-similarity, cosine similarity

For the task of ontology learning [12] as well as classification, e.g. k-nearest-neighbors, the concept of similarity is of importance. A basic intuition of similarity is for example given by Lin [15]. Similarity is related to the commonalities and differences between two objects. More commonalities implies higher similarity. Vice versa, more differences implies lower similarity. Two identical objects should have the maximum similarity. In addition, only identical objects should be able to achieve maximum similarity. Typically, similarity can be defined as a binary function, which maps two objects to a value in the interval $[0, 1]$. A value of 1 represents identical input objects. For this thesis, semantic and vector similarity measures will be used. Vector similarity.
Semantic similarity measures are needed when comparing structures, which cannot be sufficiently represented as vectors. These are for example words and classes in ontologies **citations needed**. Rodríguez and Egenhofer [19] develops a semantic similarity measure for comparing entity classes in ontologies. Entity

2.4 Similarity-based classification

Chen et al. [2]
Zhang and Zhou [22]

Explain how kNN works. Nearest-neighbors classification is a lazy method, as it does not require training before testing. This is useful for applications with high

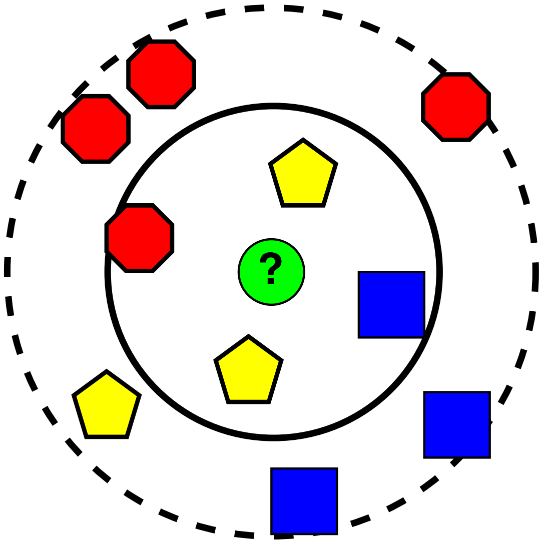

Figure 1: Example for k-nearest neighbors for 3 classes with k=4 and k=10.

amounts of data, large numbers of classes, and changing data [22]. For the considered use case of classification in Wikidata, these are very important strengths, as the number of classes in the taxonomy is very high and Wikidata is being constantly edited.

2.5 Text processing

- N-Gram
  Jurafsky and Martin [13]
- Skip-Gram
  Guthrie et al. [11]
- Counting-based word representations
  Levy et al. [14]
- Predictive word representations
  Levy et al. [14]

3 Analysis of the Wikidata taxonomy

4 Ontology learning

General concepts. Classification of considered problem in the task of ontology learning. Related work.
Cimiano et al. [3]

Wong et al. [21]
d'Amato et al. [5]
Petrucci et al. [17]
Fu et al. [8]

5 Neural networks

Notion of neural networks will be introduced.

5.1 Recursive neural networks for graph representation

Scarselli et al. [20]

5.2 Deep neural networks for graph representation

Cao et al. [1]
Raghu et al. [18]

5.3 Continuous Bag-of-Words

Mikolov et al. [16]

5.4 Skip-gram with negative sampling

Mikolov et al. [16]
Levy et al. [14]
Goldberg and Levy [10]

5.5 Comparison

6 Algorithm

6.1 Baseline

- Hyper parameters
- Training data

6.2 Supplementing with other resources

e.g. Wikipedia

7 Evaluation

7.1 Method

Dellschaft and Staab [6]

7.2 Generation of gold standard

7.3 Results

References

[1] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In Dale Schuurmans and Michael P. Wellman, editors, AAAI, pages 1145–1152. AAAI Press, 2016. URL `http://dblp.uni-trier.de/db/conf/aaai/aaai2016.html#CaoLX16`.

[2] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. J. Mach. Learn. Res., 10:747–776, June 2009. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1577069.1577096`.

[3] P. Cimiano, A. Mädche, S. Staab, and J. Völker. Ontology learning. In S. Staab and R. Studer, editors, Handbook on Ontologies, International Handbooks on Information Systems, pages 245–267. Springer, 2nd revised edition edition, 2009. URL `http://www.uni-koblenz.de/~staab/Research/Publications/2009/handbookEdition2/ontology-learning-handbook2.pdf`.

[4] Philipp Cimiano. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387306323.

[5] Claudia d'Amato, Steffen Staab, Andrea G. B. Tettamanzi, Tran Duc Minh, and Fabien L. Gandon. Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases. In Sascha Ossowski, editor, SAC, pages 333–338. ACM, 2016. ISBN 978-1-4503-3739-7. URL `http://dblp.uni-trier.de/db/conf/sac/sac2016.html#dAmatoSTMG16`.

[6] Klaas Dellschaft and Steffen Staab. On how to perform a gold standard based evaluation of ontology learning. In Proceedings of the 5th International Conference on The Semantic Web, ISWC'06, pages 228–241, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-49029-9, 978-3-540-49029-6. doi: 10.1007/11926078_17. URL `http://dx.doi.org/10.1007/11926078_17`.

[7] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to map between ontologies on the semantic web. In Proceedings of the 11th

International Conference on World Wide Web, WWW '02, pages 662–673, New York, NY, USA, 2002. ACM. ISBN 1-58113-449-5. doi: 10.1145/511446.511532. URL http://doi.acm.org/10.1145/511446.511532.

[8] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning Semantic Hierarchies via Word Embeddings. Acl, pages 1199–1209, 2014.

[9] Luis Galárraga. Rule Mining in Knowledge Bases. PhD thesis, Telecom ParisTech, 2016.

[10] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. CoRR, abs/1402.3722, 2014. URL http://arxiv.org/abs/1402.3722.

[11] David Guthrie, Ben Allison, W. Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy, 2006.

[12] Maryam Hazman, Samhaa R El-Beltagy, and Ahmed Rafea. A Survey of Ontology Learning Approaches. International Journal of Computer Applications, 22(9):975–8887, 2011.

[13] Daniel Jurafsky and James H. Martin. N-Grams. In Speech and Language Processing, chapter N-Grams. 2014.

[14] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3:211–225, 2015. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/570.

[15] Dekang Lin. An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657297.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013. URL http://arxiv.org/abs/1301.3781.

[17] Giulio Petrucci, Chiara Ghidini, and Marco Rospocher. Using recurrent neural network for learning expressive ontologies. CoRR, abs/1607.04110, 2016. URL http://arxiv.org/abs/1607.04110.

[18] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. ArXiv e-prints, June 2016.

[19] M. Andrea Rodríguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. IEEE Trans. on Knowl. and Data Eng., 15(2):442–456, February 2003. ISSN 1041-4347. doi: 10.1109/TKDE.2003.1185844. URL `http://dx.doi.org/10.1109/TKDE.2003.1185844`.

[20] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini. The Graph Neural Network Model. IEEE Transactions on Neural Networks, 20(1):61–80, jan 2009. ISSN 1045-9227. doi: 10.1109/TNN.2008.2005605. URL `http://ieeexplore.ieee.org/document/4700287/`.

[21] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Ontology learning from text: A look back and into the future. ACM Comput. Surv., 44(4):20:1–20:36, September 2012. ISSN 0360-0300. doi: 10.1145/2333112.2333115. URL `http://doi.acm.org/10.1145/2333112.2333115`.

[22] Min-Ling Zhang and Zhi-Hua Zhou. A k-Nearest Neighbor Based Algorithm for Multi-label Classification. volume 2, pages 718–721 Vol. 2. The IEEE Computational Intelligence Society, 2005. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1547385`.