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

List of Figures

1 Introduction

Motivation. Related work. Solution. Evaluation.

2 Foundations

2.1 Wikidata

TODO: Define entity in Wikidata, how are classes identified, etc. Galárraga [8] 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

- Ontology
  Cimiano [4]
  Galárraga [8]
- Taxonomy
  Cimiano [4]
  Galárraga [8]

- 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* $\sigma_R : R \rightarrow C^+$ *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* $\sigma_{\mathcal{A}} : \mathcal{A} \rightarrow C \times \mathcal{T}$*, 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:

$$
\begin{aligned}
&\forall x x \leq x \text{ (reflexive)}\\
&\forall x \forall y (x \leq y \wedge y \leq x \implies x = y) \text{ (anti-symmetric)}\\
&\forall x \forall y \forall z (x \leq y \wedge y \leq z \implies x \leq z) \text{ (transitive)}\\
&\forall x x \leq top \text{ (top element)}\\
&\forall x \forall y \exists z (z \geq \wedge z \geq y \wedge \forall w (w \geq x \wedge w \geq y \implies w \geq z)) \text{ (supremum)}
\end{aligned}
$$

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* $pred(v) = \{w \mid (w, v) \in E\}$ *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* $succ(v) = \{w \mid (v, w) \in E\}$ *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 $(c_1, c_2) \in S$.*

2.3 Similarity

- semantic similarity e.g. distributional similarity
  Lin [13]
  Rodríguez and Egenhofer [17]

- geometrical similarity e.g. distance based-similarity, cosine similarity

2.4 Similarity-based classification

Chen et al. [2]
Zhang and Zhou [20]

2.5 Text processing

- N-Gram
  Jurafsky and Martin [11]

- Skip-Gram
  Guthrie et al. [10]

- Counting-based word representations
  Levy et al. [12]

- Predictive word representations
  Levy et al. [12]

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. [19]
d'Amato et al. [5]
Petrucci et al. [15]
Fu et al. [7]

5 Neural networks

Notion of neural networks will be introduced.

5.1 Recursive neural networks for graph representation

Scarselli et al. [18]

5.2 Deep neural networks for graph representation

Cao et al. [1]
Raghu et al. [16]

5.3 Continuous Bag-of-Words

Mikolov et al. [14]

5.4 Skip-gram with negative sampling

Mikolov et al. [14]
Levy et al. [12]
Goldberg and Levy [9]

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

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

[9] 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`.

[10] 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.

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

[12] 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`.

[13] 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`.

[14] 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`.

[15] 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`.

[16] 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.

[17] 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`.

[18] 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/.

[19] 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.

[20] 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.