Extending taxonomic relations of root classes in the Wikidata knowledge base with neural networks

Proposal for Bachelor thesis

Alex Baier
abaier@uni-koblenz.de

November 26, 2016

1 Motivation

TODO 1: Reorder the motivation. Explain all mentioned concepts, with a short sentence. Show example for Wikidata item and property. Wikidata is a open, free and collaborative knowledge base by the Wikimedia foundation. **TODO 2: Write something about the usage and purpose of Wikidata.** As of the 7th November 2016 Wikidata offers 24,438,781 entities [2]. Entities are differentiated between items (e.g. Albert Einstein (Q937), linked data (Q515701), etc.) and properties (e.g. educated at (P69), has as part (P527), etc.). Of these 20 million entities 1,217,733 are classes, which form a taxonomy using the subclass of (P279) property. Like every other entity in the knowledge base, these classes are created by users (and bots), and are therefore incomplete and sometimes erroneous. This leads to an incomplete and erroneous taxonomy. One of the issues in the Wikidata taxonomy is the high amount of root classes, classes which have no parents and are therefore the highest level of abstraction in the taxonomy As of the 7th November 2016, 7142 root classes exist. The following are examples of root classes in Wikidata:

- Men's Junior European Volleyball Championship (Q169359)
- Women's Junior European Volleyball Championship (Q169956)
- **TODO 3: Add other root classes, which should not be root classes**

It can be seen with the first two examples, that both classes should share a parent class, which for example could be the European Volleyball Championship (Q6834). There are 355 other root classes with the property sport (P641), for which similar generalizations could be made.

TODO 4: explain how the decision is not always easy.
TODO 5: Explain why refining the taxonomy is useful and how it may help Wikidata.

At this time refining the taxonomy by generalizing existing root classes is a process done by human users. For other problems like entity typing, tools exist to support the Wikidata community (e.g Wikidata Game [1]).
A tool for suggesting possible parent classes for root classes could work similar to the mentioned Wikidata Game. To make such a tool possible, a method for finding generalizations of root classes has to be developed.
TODO 6: following sentence belongs to Related Work Other approaches already exist, which solve similar problems, see for example [9] which exploits the taxonomic structure of thesauri using the similarity of nearest neighbors to classify new words into its classes.
The exploitation of neural networks, which have been shown to be very powerful and versatile **TODO 7: Do I need a reference for this?**, is an approach, which **TODO 8: is this even true?: currently has not been evaluated** for the task of taxonomy refinement. It may be prove useful to design and evaluate such a system to show, how neural networks may benefit the area of taxonomy learning.

2 Problem statement

To define the problem following definitions are needed:
Definition 1 (Forest)**.** A forest is a set of trees. Let F be a forest, then $|F|$ is the number of trees in F.
Definition 2 (In-Tree, Root Node)**.** An in-tree is a directed, complete, acyclic graph. Additionally from each node in the in-tree a directed path to the root node can be found. It can be represented as a tuple (N, E), where N is a set of nodes of the tree and $E = \{(n_1, n_2)|n_1, n_2 \in N\}$ is a set of directed edges. A node in an in-tree is a root node , if it has no outgoing edges. There is only one root node in an in-tree. Let $T = (N, E)$ be an in-tree, then $root(T) \in N$ is the root node of T.
Definition 3 (Class)**.** A class is a tuple $(id, label, Statements, Instances, wiki)$:

- $id \in \mathbb{N}$, which is a numerical Wikidata item ID;
- $label$, which is the, to id corresponding, English label in Wikidata;
- $Statements$ is a set of statements about the class;
- $Instances \in P(\mathbb{N})$ is the set of numerical Wikidata item IDs, which are instances of the class;
- $wiki$ is the, to the class corresponding, English Wikipedia article text.

Definition 4 (Statement)**.**

Definition 5 (Class Hierarchy, Root Class)**.** A class hierarchy is an in-tree, in which the nodes are classes. The directed edges describe the subclass-of relation between the child and the parent class. The root node in a class hierarchy is called root class.

Definition 6 (Taxonomy)**.** A taxonomy is a forest of class hierarchies. The number of root classes in a taxonomy T is $|T|$.

Problem. Let W_1 be the taxonomy of Wikidata, where only labeled root classes are considered. On 7th November 2016 the following state applies $|W_1| = 5332$.
W_1 is the input for the described problem.
Let W_2 be the refined output taxonomy.
After the refinement process the following should apply: $|W_2| \ll |W_1|$. This means the number of class hierarchies in the Wikidata taxonomy is significantly reduced.

The refinement process can be reduced to the following smaller task:
For a class hierarchy $H_1 \in W_1$ find a class $c \in H_2 \in W_1 \setminus \{H_1\}$, which is an appropriate superclass for $r = root(H_1)$.
Combine $H_1 = (N_1, E_1)$ and $H_2 = (N_2, E_2)$ to a hierarchy $H = (N_1 \cup N_2, E_1 \cup E_2 \cup \{(r, c)\}$.
The output taxonomy is defined as follows: $W_2 = (\{H\} \cup W_2) \setminus \{H_1, H_2\}$. Accordingly $|W_2| = |W_1| - 1$ follows.
Repeating this smaller task will eventually yield $|W_2| \ll |W_1|$.

The problem is therefore to develop a method, which finds an appropriate superclass for a root class given the complete surrounding taxonomy.

3 Related works

TODO 9: Add literature about neural networks, ontology learning.

4 Methodology

TODO 10: Order is not correct, rewrite this. The main task of this work will be to develop a method for extending taxonomic relations of root classes based on neural networks. The task can be divided into the following subtasks:

The root classes in the knowledge base have to be identified and analyzed. This should result in a comparison of similarities and characteristics of root classes in Wikidata. The purpose of this task is to identify, which data is available and how it is structured.

Two categories of to this challenge related work has to analyzed. The first category of related work is concerned with the topic of ontology and taxonomy learning. The second category is concerned with different applications of neural networks, and ways to represent complex data for usage in neural networks. Goal of this task should be to find an appropiate mapping of Wikidata root classes to feature vectors, which can be used by neural networks.

After the mapping of classes is defined, data can be collected per hand or possibly by crawling for training neural networks. The author will create a ground truth for the collected data.

Finally different neural networks can be implemented and trained using existing libraries. The configurations of the networks will be improved by means of experimentation and literature review **TODO 11: Is "literature review" the right word?** until a small enough error in testing is achieved.

The real data, all 7142 root classes, will be applied on the best performing network(s). The results will be reviewed by the author. If it is possible a survey with the Wikidata community will be executed. A (random?) subset of the results will be presented to the community, and the participants of the survey will be asked, whether they think the generated suggestions of the network are accurate and could be entered into Wikidata. Such a survey would be really important to confirm the validity and relevance of this work.

5 Expected results

TODO 12: What should I expect? NNs are powerful, so it is likely to work, if the right data is used.

6 Time plan

References

[1] https://tools.wmflabs.org/wikidata-game/.

[2] https://tools.wmflabs.org/wikidata-todo/stats.php.

[3] 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.

[4] 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.

[5] Alexander Maedche, Viktor Pekar, and Steffen Staab. Ontology Learning Part One — on Discovering Taxonomic Relations from the Web, pages 301–319. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[6] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. IEEE Intelligent Systems, 16(2):72–79, March 2001.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.

[8] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, pages 1045–1048, 2010.

[9] Viktor Pekar and Steffen Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In Proceedings of the 19th Conference on Computational Linguistics, COLING-2002, August 24 - September 1, 2002 , Taipei, Taiwan, 2002, 2002.

[10] Giulio Petrucci, Chiara Ghidini, and Marco Rospocher. Using recurrent neural network for learning expressive ontologies. CoRR, abs/1607.04110, 2016.

[11] Alessandro Sperduti and Antonina Starita. Supervised neural networks for the classification of structures. IEEE Transactions on Neural Networks, 8(3):714–735, may 1997.

[12] 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.