Outline

Alex Baier
abaier@uni-koblenz.de

February 13, 2017

1 Introduction

- Motivation
- Related work
- Novelty/Challenges
  - application of neural word embeddings in OL
  - large number of classes
- Solution
- Results
- Overview

2 Foundations

2.1 Wikidata

- What is Wikidata?
- Items
- Statements
- Difference between class and instance

2.2 Taxonomy

- Notion of ontology and taxonomy
- Definition of taxonomy
- Unlinked vs. root classes

2.3 Similarity measures

- information-theoretical notion of similarity (commonalities, differences)
- Vector similarity: distance-based, cosine
- Semantic similarity

2.4 Problem statement

- Definition of problem
- Challenges

2.5 k-nearest-neighbors classification

- What is kNN?
- How is it suited to the problem?
- Formal description for weighted single-label kNN
- Multi-label kNN
- How can it be applied to the problem? Make Wikidata items comparable via similarity measure.

3 Neural networks

Answer question, how neural networks are suited to represent Wikidata items as vectors. This will allow it to apply similarity measures and therefore solve the problem.

3.1 Feedforward neural network with backpropagation using gradient descent

- Introduce neural network on simple model
- Use simple classification or regression problem as example

3.2 Deep neural networks for graph representation

3.3 Word2Vec

- N-grams, Skip-grams
- CBOW
- Skip-gram
- about hyperparameters

3.4 RDF2Vec

3.5 Comparison

4 Ontology learning

What is ontology learning and why is it interesting?

4.1 Process and architecture for ontology learning

4.2 Approaches for learning taxonomic relations

4.3 Ontology learning using neural networks

5 Analysis of the Wikidata taxonomy

5.1 Root taxonomy

5.2 Unlinked classes

5.3 Labeled, instantiated, unlinked classes

Find an acronym for this set

- Reasons for reducing the input set.
- Compare results to unlinked classes

6 Combined algorithm

6.1 Baseline

Skip-gram using Wikidata statement triples as input sentences + ml-kNN.

- Model of algorithm
- How are the challenges solved?
- Choice of hyperparameters

6.2 Variation: RDF2Vec

Skip-gram using graph walk to create input sentences from Wikidata + ml-kNN.

- What is added/changed?
- Model of algorithm
- Possible benefits/deficits

6.3 Variation: Wikipedia

Skip-gram using Wikidata statement triples as input sentences. Skip-gram using Wikipedia as input. Combine word embeddings of both NNs to one embedding and apply ml-kNN.

- What is added/changed?
- Model of algorithm
- Possible benefits/deficits

7 Evaluation

7.1 Method

7.2 Generation of gold standard

pick number (tbd) of random linked classes. the distribution of instances, subclasses, properties per class should similar to the repeated analysis of the observed unlinked classes. remove the subclass properties of the chosen classes. also generate a new set of training data, which reflects the changes, and train the model on this modified data.

7.3 Results

compare baseline algorithm and variation(s). do the results match with my expectations? if not why could this be? mention training and execution time of the algorithms.

8 Conclusion

what was learned? can the developed solution be used in practical application? what about future work?