UNIVERSITEIT VAN AMSTERDAM

MSc Artificial Intelligence
Master Thesis

Knowledge Generation

by
Florian Wolf
12393339

November 25, 2020

48 Credits
April 2020 - December 2020

Supervisor:
Dr Peter Bloem
Chiara Spruijt

Asessor:
Dr Paul Groth

Abstract

We generate Knowledge! [1]

1 Introduction

Here comes a beutiful introduction. Promise!

2 Background knowledge

3 Background

In this section I will present my literature research up-to-date. The included topics are either relevant as background knowledge or state-of-the-art models.

3.1 Knowledge

If we thjnk about, how we aquire knowledge, the most common way is audio-visually, e.g in form of a lecture, a movie, or book. A common way to pass on knowledge is by bringing it to text. For machines to be able to reason and work with knowledge, it needs to be transferred to a machine readable format. The most popular format is a tabular database. A newer database approach is the knowledge graph (KR) which is based on relations between entities. This lets the machine reason on this knowledge, answer complex questions and make conclusions more similar to human thinking. The format of KG is a triple consisting of a subject, a directed relation and an object. The task of converting knowledge from text to KG format is non-trivial. In this thesis, we would like to focus on the extraction of accurate knowledge. The idea behind it is that text possesses various levels of knowledge and we, as reader, are biased by our prior knowledge and the intention or drive for reading the text.

- Semantic Parsing
  The most old-school and technical approach is to tackle each sentence with NLP methods. This means tagging each word, linking references and finding relations. This will extract all possible triples. This can then be aligned with an existing database [2].
- Knowledge Graph
  Form of representing knowledge or simply a database. Makes information machine readable. It is build of relation triples.
- Triple SRO
  Triples are formed by two nodes and on link. The starting node is the subject, the second node is the object and the directed link is the relation of the subject with the object.
- Semantic Web
  The future of the internet where the data presented on websites can be read and understood by the browser. For this to be possible, websites should present its information in KG format in the metadata.
- Existing Methods in NLP
  Different semantic parser have been developed. They extract all possible kind of triples from text. While being grammatically correct, the extracted triples do not represent the information a human reader would get by reading the text. The Stanford Parser might be the most popular and advanced one [3]. Further the Allen NLP and OpenIE projects offer powerful parsing tools.

3.2 Embeddings

For any model we need an encoding to input text. Word embeddings have established themselves over the last decade as the solution. There are differences in how they are trained , how the relations between words is captured and how the context is represented.

- Word Embeddings
  word2vec is the most established word embedding, easy to train and implement [4]. Yet this method is becoming outdated and being replaced by newer solutions.

  BERT, also by Google, seems to be the sate of the art. It is able to predict words or full sentences

as vectors. Using a bidirectional architecture and an attention score for each token, this model is able to catch much more context than its predecessors [5]. The attention score indicates how much other words point towards the selected word within a sentence [6]. This way the importance of the words can be compared. The model on it's own is a text classifier but can be tweaked to output word embeddings.

- Graph Embeddings
  TransE represents entities in in low-dimensional embedding. The relationships between entities are represented by the vector between two entities [7]. (How are different relation between the same entities represented?)

- OntoUSP
  This method learns a hierarchical structure to better represent the relations between entities in embedding space.

3.3 Learning Methods

A big challenge of this field is the lack of labeled data. Public knowledge graphs like DBpedia hold huge collections of knowledge. In my opinion, the size is the problem. The plain amount of data makes it hard to incorporate it in an efficient pipeline. Further, since it needs to cover all topics, the representation becomes less specific. Looking at the usecases of KGs, most tasks actually do not require this total coverage and instead are topic specific. Thus, we would like to focus on a targeted approach.

- Supervised Learning
  Requires a labeled dataset. For the case of text to graph the options are limited. Adecorpus provides such a dataset for drug interactions. Facebook research offers a QA text to graph dataset called babi. This dataset is a benchmark for question answering algorithms.

- Distant supervision
  makes use of the a database like DBpedia and infers labels by comparing the similarity of entities.

- Contrastive Learning
  Can be supervised or unsupervised approach and focuses on similarities between predictions. The loss is computed by the energy function of the output. To me this seems like an interesting approach to create a world model [1] and it has not yet been applied to plain text.

3.4 Recurrent Graph Models

The idea of creating a graph recurrently node by node seems intuitive. One recent example of generating a graph using a recurrent VAE architecture was [8]. Here the model generates gets bird-view images of roadmaps and generates a graph representation. each generated node is fed back into the encoder as prior and a stop signal is generated once finished. Applying this to knowledge from text is one of my two main ideas 8.2.

Another recent approach uses recurrent Graph GAN [9]. Learns distribution of the training graphs. Creates Graph sequentially.

3.5 Graph Normalizing Flows

Create Graph all at once. Would it be possible to generate it recursively? Conditioned on query?

- unsupervised learning with CFG In the application of context free grammar, NFs have been trained on plain text [10]. The approach was unsupervised with a loss function which takes in account the distribution of the output prediction and the KL divergence between the distributions of each output. This made the model generate outputs with high certainty and which lay far apart from each other. The loss also encourages the output to be close to the input. For the input and output two different embeddings are needed. Thus, we need a similarity measure between them.

3.6 Variational Auto Encoder (VAE)

These models consist of an encode and a decoder. The encoder encodes the input to an low-dimensional latent space. The decoder takes a signal from latent space and reconstructs the input. Exploring the latent space makes it oossible to use the decoder as generative model. The posterior can only be approximated by the ELBO

4 Related Work

In this section we will go over related work and relevant background information. The depth of the explanation is adopted to the expected prior knowledge of the reader. The reader is supposed to know the basics of machine learning, including probability theory and basic knowledge on neural networks and their different architectures.

4.1 The Graph VAE – one shot method

VAE

The VAE as first presented by [11] is an unsupervised generative model consisting of an encoder and a decoder. The architecture of the VAE differs from a common autoencoder by having a stochastic module between encoder and decoder. Instead of directly using the output of the encoder, a distribution of the latent space is predicted from which we sample the input to the decoder. The reparameterization trick allows the model to be differentiable. By places the sampling module outside the model we get a deterministic model which can be backpropagated.

MLP

The Multi-Layer Perceptron (MLP) was one on the first machine learning models. In its basic structure it takes a one dimensional input, fully-connected hidden layer, activation function and finally output layer with normalized predictions. Images or higher dimensional tensors can be processed by flattening them to a one dimensional tensor and adjusting the input dimensions. This makes the MLP a flexible and easy to implement model.(reference)

Graph convolutions

CNNs have shown great results in the field of images classification and object detection. This is due to the fact that a convolution layer takes into account the relation of one pixel to its neighbors. The same holds for graph CNNs where at each convolution the information of each node is passed on as messages to all its neighbors. Each convolution applies an activation function in between the steps. In this case the information is the edges each node has. To process node attributes and edge atributes we have to look at more complex models [12].

RGCN

Realtional Graph Convolution Net (RGCN) was presented in [13] for edge prediction. This model takes into account features of nodes. Both the adjacency and the feature matrix are matrix-multiplied with the weight matrix and then with them-selves. The resulting vector is a classification of the nodes.

Graph VAE

Now we have all the building bocks for a Graph VAE. The encode can either be a MLP, a GCNN or an RGCN. The same holds for the decoder with the addition that model architechture needs to be inverted. An verison of a Graph VAE presented in [14]. This model combines both the previous methods. The input graph undergoes relational graph convolutions before it is flattened and projected into latent space. After applying the reparametrization trick, a simple MLP decoder is used to regenerate the graph. In addition the model concatenates the input with a target vector y, which represents ???. The same vector is concatenated with the latent tensor. ***Elborate why they do that***.

Graphs can be generated recursively or in an one-shot approach. This paper uses the second approach and generates the full graph in one go. ***Cite?***

4.2 Permutation Invariance

Permutation invariance refers to the invariance of a permutation of an object. An visual example is the the image generation of numbers. If the loss function of the model would not be permutation invariant, the generated image could show a perfect replica of the input number but due to positional permutation the loss function would penalize the model. OR: An example is in object detection in images. An object can have geometrical permutations such as translation, scale or rotation, none the less the model should be able to detect and classify it. In that case, the model is not limited by permutations and is there fore permutation invariant. In our case the object is a graph and the nodes can take different positions in the adjacency matrix. To detect similarities between graphs we apply graph matching.

Graph matching algorithms

There are various graph matching algorithms. The one we will implement is the max-pooling (Finding Matches in a Haystack: A Max-Pooling Strategy for Graph Matching in the Presence of Outliers).

The max-pooling graph matching algorithm presented by

The algorithm returns a symmetric affinity matrix for all nodes

The resulting similarity matrix gives us $X*$ which is continuous and therefore useless. To transform is to a discrete X we use the hungarian algorithm (GPU-accelerated Hungarian algorithms for the Linear Assignment Problem)

Hungarian algorithm
The hungarian algorithm is used to find the shortest path within a matrix. This could be the most efficient work-distribution in a cost matrix. Or ...
It consists of four steps of which the last two are repeated until convergence. This algorithm is not scalable. The Munks aalgorithm (refference) takles this problem by?? and is scalable.

second option: Compare only graph structure. NX algorithms: Greedy, Shortest path ...

Graph VAE loss The loss is discribed in [14] as.

In contrast to his implementation we assume, that a node or edge can have none, one or multiple attributes. Therefore our attributes are also not sigmoided and do not sum up to one. This leads o the modification of term logF and logE where we do not matrix multiply over the attribute vector but take the BCE as over the rest of the points. KG can have multiple or no attributes vs molecular graphs can be one hot encoded.

Okay, further we need to treat the $log_p E$ and $log_p F$ just like $log_p A$ and subtract the inverse. Otherwise the model learn to predict very high values only.

A note to the node features, these stay softmaxed and one-hot encoded since we will use them as node labels.

4.3 Knowledge Graphs

Knowledge Graphs are great! The best in the world.

5 Metodology

This section describes the methodology for this thesis. The first part includes the presentation of the model, the reprocessing of the input and the evaluation metrics. The second part describes the experimental setup and the different experimental runs. The work of this thesis has aimed to be fully reproducable, thus the code is opensourced and available on Github [1].

5.1 Knowledge graph representation

The first step in our pipeline is the representation of the KG in tensor format. In order to represent the graph structure we use an adjacency matrix A of shape $n \times n$ with n being the number of nodes in our graph. The edge attribute or directed relations between the nodes are represented in the matrix E of shape $n \times n \times d_E$ with d_E being the number of edge attributes. Similarly for node attributes we have the matrix F of shape $n \times d_N$ with d_N number of node attributes. The input graph can have less nodes than the maximum n but not more. The diagonal of the adjacency matrix is filled with 1 if the indexed node exists, and with 0 otherwise. The number and encoding of the attributes must be predefined and cannot be changed after training. This way we can uniquely represent a KG.

Graph embeddings? unsupervised approach

5.2 Graph VAE

hi
Convolution part
RCGN relation Convolution neural net
MLP encoder
Latent space
reparametrization trick
MLP decoder
Graph matching
Discretization of prediction

5.3 Loss function

If loss function should be permutation invariant we need to do some kind of graph matching.
Different options for graph matching.
Maxpooling-algorithm:
Assumptions
Node to edge affinity equals 0

[1] ***Thesis Repo Link***

Self-loops are possible, adjacency matrix can be zero or one.

Summing over the neighbors means summing over the whole column Normalize matrix with Frobenius Norm
Batch version: Only matmul and dot. keep dimension of S with shape (bs,n,n,k,k) When maxpooling, flatten Xs (n,k) for batch dot multiplication. This way (i think) we sum over all j nad b neighbors instead of taking the max.
Hungarian algorithm for discrimination of X

The hungarian algorithm as presented in section 2 return the shortest path in a matrix. We use this shortest path as bast match between the two graphs. The node paris identified as optimal are masked as 1 and the rest of the matrix as 0. This way we discretizice $X\star$ to X.
The equations

6 Experiments

This section presents the experiments we ran. We covered link and node prediction and compared those to SOTA scores. Further we ran experiments on investigating the coherence of the reproduced graph structure. Lastly we measured the adherence of our model to the KG's underlying syntax.

6.1 Datasets

For this sake of meaningful results, we chose to use the two most popular dataset used in this field.

FB15k237 A subset from on of the first and largest KGs, called FreeBase. In the first version of this dataset, it was possible to infer most of the test triples by inverting trples in the trainset, Thus the latest 237 version filtered these triples out. The dataset contains $14,951$ entities and $1,345$ different relations.

WN18kRR Subset from the KG WordNet.

6.2 Link Prediction

We used link prediction as evaluation protocol and for comparison state of the art models. For each triple in the dataset, we remove the tail and combine it with all possible entities in our dataset. EVen though this is called 'triple-corruption', also correct triples can be generated, which could appear in the trainset. These have bo the filtered out before evaluation. Link prediction on unfiltered test data is termed 'raw'. Our model then computes the ELBO loss for all corrupted triples, which are sorted and stored in ascending order. The same procedure is repeated the triple's head in place of the tail.

The metrics used to evaluate the predictions, is the mean reciprocal rank (MRR) and hits at 1,3 and 10. The MRR ??? Hits at 1 indicates what percentage of the triples in the test set have been ranked the highest compared to their corruptions. Similar, hits at 3 and 10, give a percentage of triples ranked in the top 3 and top 10. These metrics allow a fair comparison on a dataset between models, regardless of the ranking method of each model.

6.3 Sanity Checks

6.4 Syntax coherence

7 Results

8 Discussion & Future Work

Ideas

The bigger picture of this thesis is to efficiently generate a representation of the information hold in plain text. This representation has the form of a knowledge graph and consists of subject-relation-object triples.
Challenges:

- What knowledge are we looking for? I would like the model to focus on the most important information. This could for example be topic specific.
- Another option would be to extract based on a query or point of interest. Especially if we build the KG incrementally we can use this input as starting point and recursively build from there.

8.1 Use Normalizing Flows

- Is it possible to generate a graph from text using Normalizing flow (NF)?
- Can we train such a NF unsupervised, using the output distributions as loss function?
- NFs can build KG at once or incrementally. No one is using it, could have a reason.

8.2 Recurrent VAE

- Recursively generate graph using a Variational Auto Encoder.
- Use query as start node.
- expand graph on that node. Thiviyan and others work on VAEs as well. Seems an easier approach. Can analyze the latent space.

The encoder decoder strategy has been applied successfully in many cases. Bellis thesis about modeling a graph from images can be applied to text as well by changing the input to a text vector [8]. Note that this is a supervised method and would require a dataset of text labeld with resulting KGs. To the best of my knowledge this does not exist yet.
Same holds for the contrastive world model by Kipf. If we input text vectors instead of an image the model could recognize different objects in the text as it does with pixels [1]. Here the model is trained unsupervised using a loss over the energy function of the graph embedding space TransE [7]. An open question is if this would work for our approach.
A bit more abstract is the idea of the feedback recurrent VAE [15] where sound signals are encoded and decoded. This could also be adopted to text for instance with one sentence at a time, or a fixed number of tokens. Here the text vector would be induced as the latent input to the decoder. This would mean finding an translation of the models latent space to the work embedding. The dataset would consist of positive and negative examples of resulting graphs over timesteps.
While text can be easily vectorized by word embeddings like word2vec, the graph representation seems more tricky. A reasonable approach following Bellis example would be to output a coordinates vector for the nodes and an adjantency matrix for their relationships. Here one node at a time is outputted and the relation is conditioned on the number of previous nodes. Alternatively we could make use of the graph embeddings RASCAL or TransE. Lastly the question remains how to model the graph when the nodes need to be predefined. A subgraph of DBpedia could be a good starting point.

Open Issues

A section where note will be made on open issus. These aissues are ment to be discussed with Peter or Thivyian and ultimatly solved. No open issues should remain at the end of November.

8.3 Graph matching

Hungarian algorithm: Should we assume the affinity matrix is a cost or a profit matrix. If we assume it is a cost matrix and n != k the algorithm returns an error while padding. If we assume it is a profit matrix and convert it to a profit matrix it can handle any dimension. The shortest path varies with both approaches. Further I assume the shortest path is what we are looking for and mask these entries with 1 and the rest of the matrix with 0.
Diagonal of A is zeros. As soon as I do this, code crashes. Think it is because of divided zero, Martin adds 1e-7 to the X norm
Do we backpropagate through the graph matching? then all torch functions, otherwise np.
Martin takes 300 iterations for hungarian. also not looping over pairs.

References

1. Kipf, T., van der Pol, E. & Welling, M. Contrastive Learning of Structured World Models. *arXiv:1911.12247 [cs, stat].* arXiv: 1911.12247. `http://arxiv.org/abs/1911.12247` (2020) (Jan. 2020).
2. Kertkeidkachorn, N. & Ichise, R. *T2KG: A demonstration of knowledge graph population from text and its challenges* in. **2293** (2018), 110–113.

3. Che, W., Liu, Y., Wang, Y., Zheng, B. & Liu, T. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. *arXiv:1807.03121 [cs]*. arXiv: 1807.03121. http://arxiv.org/abs/1807.03121 (2020) (July 2018).
4. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. arXiv: 1301.3781. http://arxiv.org/abs/1301.3781 (2020) (Sept. 2013).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805 (2020) (May 2019).
6. Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]*. arXiv: 1706.03762. http://arxiv.org/abs/1706.03762 (2020) (Dec. 2017).
7. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. in *Advances in Neural Information Processing Systems 26* (eds Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 2787–2795 (Curran Associates, Inc., 2013). http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf (2020).
8. Belli, D. & Kipf, T. Image-Conditioned Graph Generation for Road Network Extraction. *arXiv:1910.14388 [cs, stat]*. arXiv: 1910.14388. http://arxiv.org/abs/1910.14388 (2020) (Oct. 2019).
9. Li, Y., Vinyals, O., Dyer, C., Pascanu, R. & Battaglia, P. Learning Deep Generative Models of Graphs. *arXiv:1803.03324 [cs, stat]*. arXiv: 1803.03324. http://arxiv.org/abs/1803.03324 (2020) (Mar. 2018).
10. Jin, L., Doshi-Velez, F., Miller, T., Schwartz, L. & Schuler, W. *Unsupervised Learning of PCFGs with Normalizing Flow* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, July 2019), 2442–2452. https://www.aclweb.org/anthology/P19-1234 (2020).
11. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*. arXiv: 1312.6114. http://arxiv.org/abs/1312.6114 (2020) (May 2014).
12. Tiao, L., Elinas, P., Nguyen, H. & Bonilla, E. V. Variational Graph Convolutional Networks. en, 11.
13. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. en. *arXiv:1609.02907 [cs, stat]*. arXiv: 1609.02907. http://arxiv.org/abs/1609.02907 (2020) (Feb. 2017).
14. Simonovsky, M. & Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv:1802.03480 [cs]*. arXiv: 1802.03480. http://arxiv.org/abs/1802.03480 (2020) (Feb. 2018).
15. Yang, Y., Sautière, G., Ryu, J. J. & Cohen, T. S. Feedback Recurrent AutoEncoder. *arXiv:1911.04018 [cs, eess, stat]*. arXiv: 1911.04018. http://arxiv.org/abs/1911.04018 (2020) (Feb. 2020).