Master Thesis

Kavita Chopra

presented on 21.09.2016

Superviser: Denis Lukovnikov

Topic:

Prediction of the next entity/relation in a sequence of triples derived from Knowledge Graphs using Recurrent Neural Networks

(context: question answering using web data)

Content

- Knowledge Graph Modeling
- Neural Networks
- Roadmap to 

Knowledge Bases

- Structure the web to draw semantic potential for machine readability
- Store knowledge in triple formats
- Increasing in size and number
- Expansion through human effort and automated extraction from structured and non-structured sources

Knowledge Bases

- *“Expansion through human effort and automated extraction from structured and non-structured sources“*

KBs are often incomplete and might not be error-free!

→ need of **quality control**:

- Motivates statistical modeling of Knowledge Graphs for *Knowledge Base Completion* (KBC) and cleansing tasks

Facts and Figures

- KB are on the rise
- Initially driven by academic efforts, e.g. former Freebase, Yago
- ...more recently commercial applications:
  - In 2010: Freebase bought by Google and is now powering Google Knowledge Graph and Google Knowledge Vault which supports search engine and other Google applications, e.g. Google Now
  - Other applications relying on KGs: Bing, Cortana (Microsoft Satori), IBM Watson

Facts and Figures

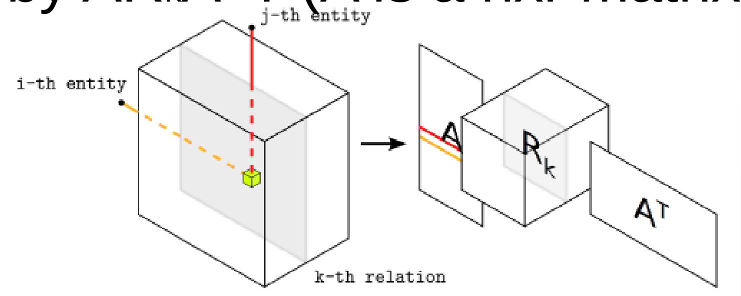
- Missing facts in KBs, e.g.:
  - in both Freebase and DBPedia ~70% of the persons are missing a place of birth
  - in DBPedia ~60% of the scientist do not have a fact describing what they are known for
  - Also in DBPedia, 40% of all countries miss a capital and 80% of the capitals do not have a leader
  - Generally, the amount of missing information is even higher for less popular entities and relations
  - Heavy-tailed distribution: the majority of the triples are about a small set of entities

Knowledge Base Completion

- Derive new facts out of existing facts
- Identify inconsistency due to false facts
- Requires representation learning
  - Latent variable embedding for entities and relations providing better representations of their semantic relationships

Latent Variable Models for KB

- State of the art models:
- TransE (Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in Neural Information Processing Systems. 2013.)
  - Energy based model which models relations between entities as translations in the embedding space
  - Confidence into a fact: similarity of the translation of the subject embedding to the object embedding: dist(s+l,o)
- Rescal (Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel. "A three-way model fot modlr collective learning on multi-relational data." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.)
  - Factorization of a 3-way-tensor (X_{nxnxm}), such that each slice X_k approximated by AR_kA^T (A is a nxr matrix, R_k is a rxr matrix)



Neural Networks

- Around since the 40s (Hebb, Pitts and McCulloch)
- Hypes and downs in-between, never really took off until recently:
- Since 2010 NN have been revolutionizing the states of records in many areas
- Promising results for various tasks in NLP, topic classification, sentiment analysis, question answering and language translation
- The factors that helped NN resurge and go deep:
  - Moslty: better machine performances and better GPUs
- Recurrent Neural Networks (RNNs) are the method of choice for prediction from sequences
  - Versatile: accepts sequences as input and/or output
  - Challenging to train

Roadmap

- Extensive literature research
- Learn Tensorflow
- Implementation of KBC models:
  - TransE
  - Rescal
  - …

- RNN
  - Create training and test data, e.g. using random walks
  - Configure RNN for sequence to singular value prediction for the given setting
  - Cope with RNN specific challenges


THANK
YOU