Information Retrieval and Knowledge Graphs. The Semantic Web Technologies

I. Makarov & L.E. Zhukov

BigData Academy MADE from Mail.ru Group

Network Science

Lecture outline

1. Knowledge Graph
2. KG Retrieval from NL Texts
   - KG Completion
   - KG Reasoning
   - KG Applications

Knowledge Graphs

Main idea

- Knowledge as graphs (linked data)
- Nodes as entities
- Labels as attributes
- Edges as relation types (heterogeneous network)

Applications

- Analytic representation of data
- Interpretable decision making
- Reasoning & QA
- Edges as relation types (heterogeneous network)

The Semantic Web

RDF representation

- r(s,p,o)= subject–predicate-object relation
- ABox representing data
- TBox representing rules (ontologies)
- rdfs:domain, rdfs:range, rdf:type, rdfs:subClassOf, rdfs:subPropertyOf
- owl:inverseOf, owl:TransitiveProperty, owl:FunctionalProperty

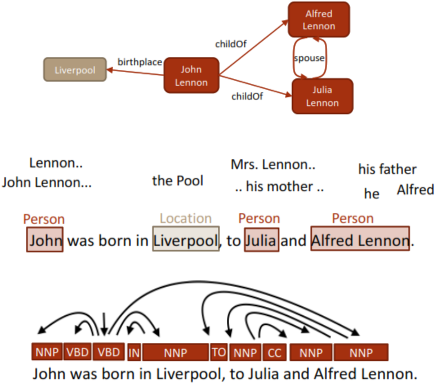
Knowledge Extraction vs. KG Construction

Problem	KE	KG
Who are entities?	NER & Coreference	Entity Linking
What are the attributes?	NER	Classification
How are they connected?	Relation extraction	Link Prediction

Table: Difference in view on knowledge mining

Knowledge Extraction

- Entity resolution, Entity linking, Relation Extraction (corpora)
- Coreference resolution (document)
- Dependency parsing, part of speech tagging, NER (sentence)



from https://kgtutorial.github.io/, 2018

Knowledge Extraction Methods

- Tagging parts of speech: CRF, CNN, bi-LSTM
- Detecting and classifying names: rules, vocabulary, DL
- Relations by dependency patterns + pronouns coreference
- Entity linking by candidate generation via entity coherence and neglecting by entity type
- Dependency parsing, part of speech tagging, NER (sentence)

Human in the loop

- Define domain (vocabulary, taxonomy, ontology)
- Learn extractor
- Score facts
- Manual — semi-automated — automated
- Human Efforts & Precision vs. Speed & Recall

Automating Knowledge Extraction

Domain

- Human made
- Partial labelling and transfer learning for semi-supervised detection
- Any noun and verb are candidates

Extractor

- Manual labelling
- Templates and manual post-processing
- Cluster SVO patterns by NER types

Scoring

- Manual scoring
- Learning scoring over labelled and unlabelled data
- Support and confidence metrics for extracted patterns compared to all the detected patterns

Knowledge Extraction Systems

	Domain	**Extractor**	**Scoring**	Fusion
ConceptNet	Human	Human	Human	
NELL	Human	Semi-Automated	Automated	Heuristics
Knowledge vault	Automated	Automated	Semi-Automated	Classifier
OpenIE	Automated	Automated	Semi-Automated	

Table: Knowledge Extraction Systems

from https://kgtutorial.github.io/, 2018

Knowledge Extraction Problems

- ambiguity
- incompleteness
- inconsistency

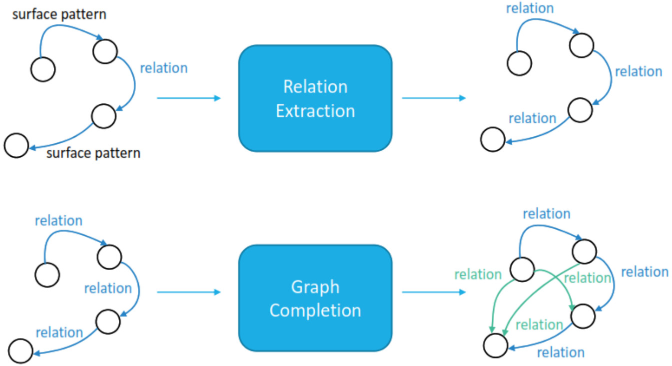
Solutions:

- Probabilistic reasoning & rule mining
- Random walks and personalized PR
- Proof construction for reasoning over KG
- Pair of nodes and relation embedding

from https://kgtutorial.github.io/, 2018

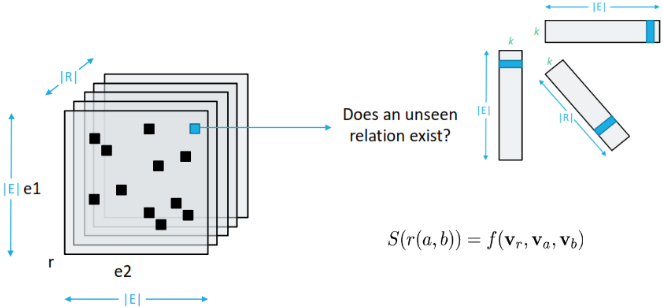
Relation Extraction and KG Completion

- Similar Pairs of Entities refer to similar relations (not identical)
- Similar Relations refer to paraphrases or implications
- Logical rules → Embedding space



from `https://kgtutorial.github.io/`, 2018

Tensor Formulation of KG



$$S(r(a,b)) = f(\mathbf{v}_r, \mathbf{v}_a, \mathbf{v}_b)$$

from https://kgtutorial.github.io/, 2018

KG Completion

CANDECOMP/PARAFAC-Decomposition

$$S\left(r(a,b)\right) = \sum_k R_{r,k} \cdot e_{a,k} \cdot e_{b,k}$$

Tucker2 and RESCAL Decompositions

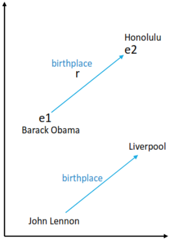
$$S\left(r(a,b)\right) = \left(\mathbf{R}_r \times \mathbf{e}_a\right) \times \mathbf{e}_b$$

Model E

$$S\left(r(a,b)\right) = \mathbf{R}_{r,1} \cdot \mathbf{e}_a + \mathbf{R}_{r,2} \cdot \mathbf{e}_b$$

Holographic Embeddings

$$S(r(a,b)) = \mathbf{R}_r \times (\mathbf{e}_a \star \mathbf{e}_b)$$



TransE

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a + \mathbf{R}_r - \mathbf{e}_b\|_2^2$$

TransH

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a^{\perp} + \mathbf{R}_r - \mathbf{e}_b^{\perp}\|_2^2$$

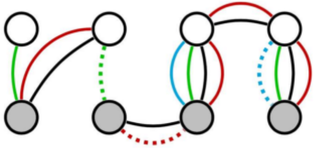
$$\mathbf{e}_a^{\perp} = \mathbf{e}_a - \mathbf{w}_r^T \mathbf{e}_a \mathbf{w}_r$$

TransR

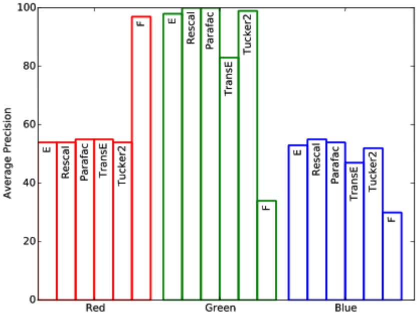
$$S\left(r(a,b)\right) = -\|\mathbf{e}_a \mathbf{M}_r + \mathbf{R}_r - \mathbf{e}_b \mathbf{M}_r\|_2^2$$

from `https://kgtutorial.github.io/`, 2018

KG Completion



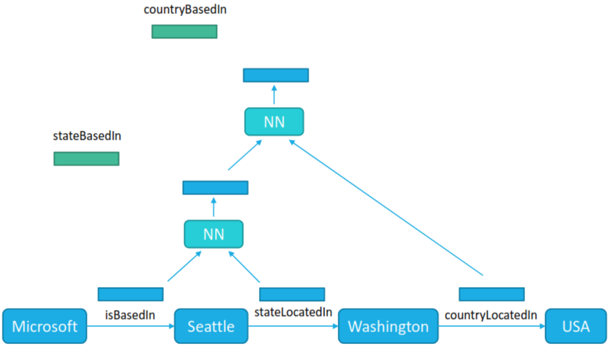
- **Red**: deterministically implied by **Black**
  - needs *pair-specific* embedding
  - Only **F** is able to generalize
- **Green**: needs to estimate entity types
  - needs *entity-specific* embedding
  - Tensor factorization generalizes, **F** doesn't
- **Blue**: implied by **Red** and **Green**
  - Nothing works much better than random



from Singh et al., 2015

KG Completion

- Compose different relations over consequent embedding from texts
- Use Neural Networks instead of Reasoning
- Construct hierarchy in automated way merging pairs and relations based on task-dependent scoring from DL model



from Singh et al., 2015

KG Challenges

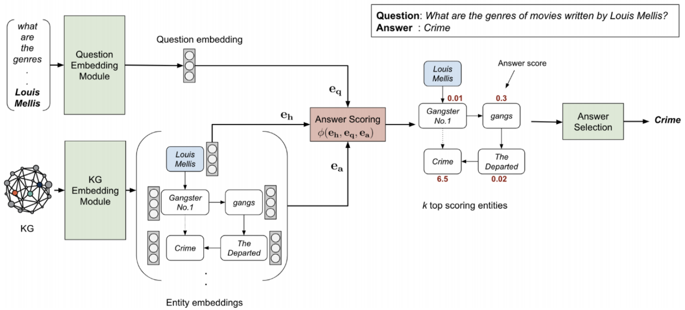
- Multi-language generalization
- Dealing with multi-modal data
- Visual-aware KG construction & captioning
- Temporal construction, correction, justifying
- Dealing with specific entities (dates, slang words)
- Changing semantics over time and language evolution

Applications

https://towardsdatascience.com/knowledge-graphs-in-natural-language-processing-acl-2020

KG: QA

- ComplEx embedding of KG, RoBERTa for question embedding
- Triple (main entity in question, question, answer in 2-hop neighborhood of main entity)
- Use Neural Networks instead of Reasoning
- Construct hierarchy in automated way merging pairs and relations based on task-dependent scoring from DL model



from Saxena et al., 2020

KG: QA

- NeuInfer architecture
- Hierarchy Mixing

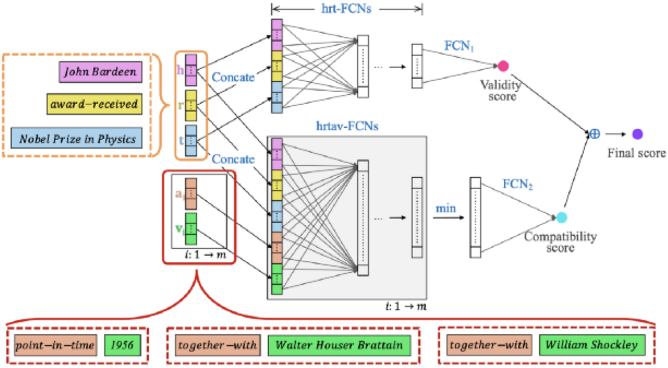


Figure 1: The framework of the proposed NeuInfer method.

from Guan et al., 2020

- Transformers rule out !

Input Table

Attribute (R)	Value (V)
Name	Walter Extra
Nationality	German
Occupation	Aircraft designer and manufacturer
...	...

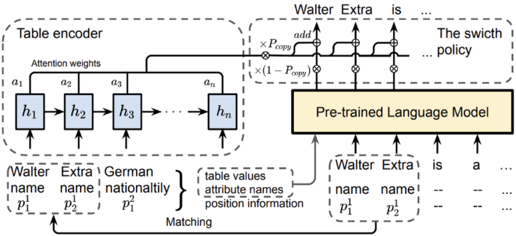


Figure 1: Overview of our approach: Under the base framework with switch policy, the pre-trained language model serves as the generator. We follow the same encoder as in (Liu et al., 2018). The architecture is simple in terms of both implementation and parameter space that needs to be learned from scratch, which should not be large given the few-shot learning setting.

from Chen et al., 2020

KG-to-Text

- Graph $\rightarrow$ Text $\rightarrow$ Graph(s) generation

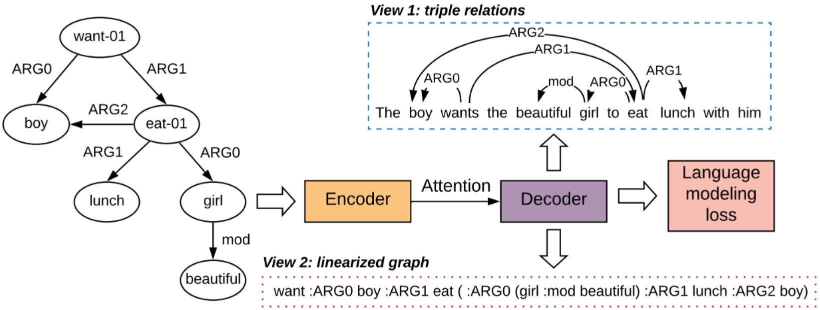


Figure 2: The training framework using multi-view autoencoding losses.

from Song et al., 2020

KG-to-Text

- R-GCN for embeddings of bi-gram relations from triple s-p-o
- Planner for counting used relations
- LSTM Decoder

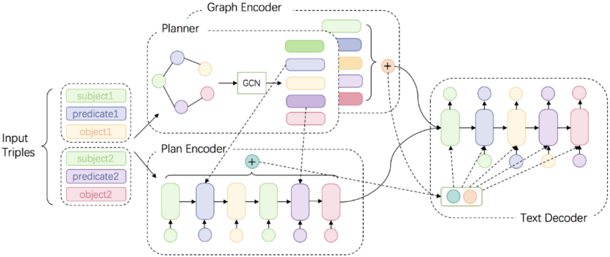


Figure 2: The architecture of the proposed DUALENC model. The input triples are converted as a graph and then fed to two GCN encoders for plan and text generation (Planner and Graph Encoder, top center). The plan is then encoded by an LSTM network (Plan Encoder, bottom center). Finally an LSTM decoder combines the hidden states from both the encoders to generate the text (Text Decoder, middle right).

from Zhao et al., 2020

KG-to-Text

- Attention from Transformer and GAT on OpenIE
- Training using RL on extracting OpenIE graphs from human-written summaries and generating questions — QA model inside !
- GPT-3 idea — train what you can

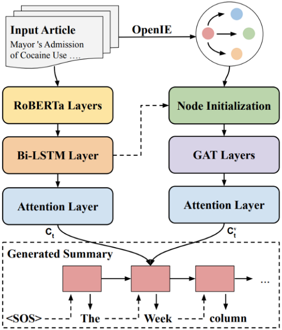
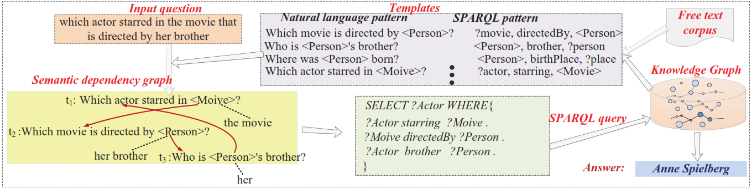


Figure 2: Our ASGARD framework with document-level graph encoding. Summary is generated by attending to both the graph and the input document.

from Huang et al., 2020

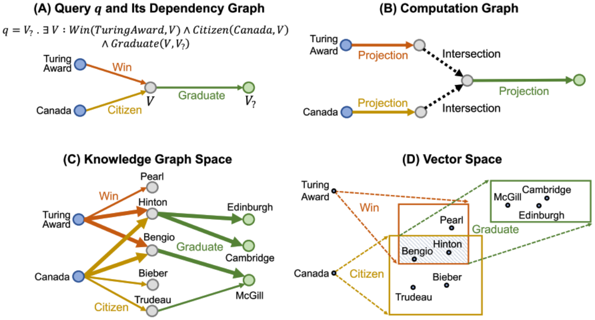
Neural, Symbolic and Neural-Symbolic Reasoning on KGs

- Neural Reasoning as Logic Query Embedding
- Symbolic Reasoning as
- Combined approach tends to extract graph and quantify its usability to the task



Zhang et al., 2021

- Use disjunctive normal form
- Consider conjunctive parts separately
- Box projection and intersection for unifying results
- Extension in BetaE for negations



Leskovec et al., 2018

- Embed paths in User-Item-Entity
- Extract Similarities via KG-based Embedding instead of User-Item decomposition

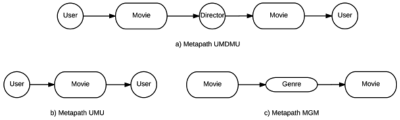


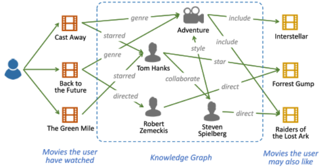
Figure 2: Movie Network Meta Paths

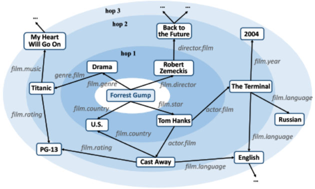
Table 1: Metapaths captured from IMDB schema

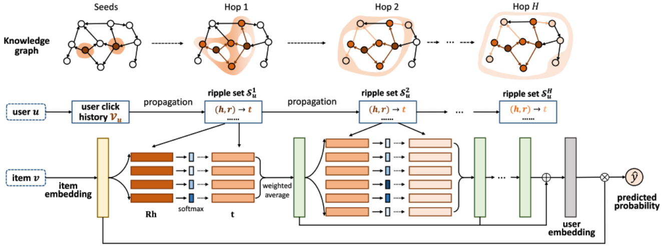
IMDB Metapaths
user - movie
user - movie - director - movie
user - movie - actor - movie
user - movie - genre - movie
user - movie - language - movie
user - movie - keyword - movie
movie - genre - movie - director
director - movie - actor - movie
director - movie - genre - movie
language - movie
keyword - movie

from Kallumadi, Surya, and William H. Hsu., 2018

KG-to-RecSys







from Guo et al., 2018

Open Challenges

- Leaks in Evaluation, Negative Sampling
- Tensor Decomposition for small KG
- Extracting n-ary relations
- Integration in IR is hard if KG quality is poor
- Reasoning/ontology always face complexity issues

Tutorials

- `https://sites.google.com/site/knowxtext/`
- `https://neo4j.com/developer/graph-data-science/build-knowledge-graph-nlp-ontologies/`
- `https://dzone.com/articles/text-mined-knowledge-graphs-beyond-text-mining`

Reasoning over ontology:

- TBox: "$Male \vee Female \rightarrow Human$"
- Boolean Query: "$Male(x) \wedge Knows(x, y) \wedge Female(y)$"
- L/NL complexity → Theoretical Computer Science

References

- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems, 2787–2795. 2013.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint, 2014.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In International Conference on Machine Learning, 2071–2080. 2016.
- Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. Holographic embeddings of knowledge graphs. In AAAI, 1955–1961. 2016.
- Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In Procs of AAAI. 2018.

References

- Stanovsky, Gabriel, Julian Michael, Luke Zettlemoyer, and Ido Dagan. "Supervised open information extraction." In Proceedings of the 2018 NACL, pp. 885-895. 2018.
- Xiong, Wenhan, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. "One-shot relational learning for knowledge graphs." arXiv preprint arXiv:1808.09040. 2018.
- Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. "COMET: Commonsense transformers for automatic knowledge graph construction." arXiv preprint arXiv:1906.05317. 2019.
- Liu, Zhibin, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. "Knowledge aware conversation generation with explainable reasoning over augmented graphs." arXiv preprint arXiv:1903.10245. 2019.
- Al-Moslmi, Tareq, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. "Named entity extraction for knowledge graphs: A literature overview." IEEE Access 8, p. 32862-32881. 2020.

References

- Qian, Wei, Cong Fu, Yu Zhu, Deng Cai, and Xiaofei He. "Translating Embeddings for Knowledge Graph Completion with Relation Attention Mechanism." In IJCAI, pp. 4286-4292. 2018.
- Kallumadi, Surya, and William H. Hsu. "Interactive Recommendations by Combining User-Item Preferences with Linked Open Data." In Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 121-125. 2018.
- Wang, Hongwei, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. "Ripplenet: Propagating user preferences on the knowledge graph for recommender systems." In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 417-426. 2018.
- Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. Proc. of NACL, 327–333. 2018.