Neural Variational Entity Set Expansion for Automatically Populated Knowledge Graphs

Pushpendre Rastogi[°][*] Adam Poliak[°] Vince Lyzinski[△][†] Benjamin Van Durme[°]

[°] Johns Hopkins University [△] UMass. Amherst

Abstract

We propose Neural Variational Set Expansion to extract actionable information from a noisy knowledge graph (KG) and propose a general approach for increasing the interpretability of recommendation systems. We demonstrate the usefulness of applying a variational autoencoder to the Entity Set Expansion task based on a realistic automatically generated KG.

1 Neural Variational Set Expansion

Imagine a physician trying to pin-point a specific diagnosis or a journalist investigating abuses of governmental power. In both scenarios, a *domain expert* may try to find answers based on prior known, relevant entities – either a list of diagnoses of with similar symptoms that a patient is experiencing or a list of known conspirators. Instead of manually looking for connections between potential answers and prior knowledge, a *searcher* would like to rely on an automatic *Recommender* to find the connections and answers for them, i.e. related entities.

In the information retrieval (IR) community, Entity Set Expansion (ESE) is the established task of recommending entities that are similar to a provided seed of entities. The physician and journalist in our example can not fully take advantage of IR advances in ESE for two main reasons. Recent advances 1) often assume access to a clean, large Knowledge Graph and 2) are uninterpretable.

Many advanced ESE algorithms rely on manually curated, clean Knowledge Graphs (KG). In real-world settings, users rarely have access to clean KGs, and instead may rely on automatically generated KGs. Such KGs are often *noisy* because they are created from complicated and error-prone NLP processes. For example, automatic KGs may include duplicate entities, associations (relations) between entities may be missing, and entities with similar names may be incorrectly disambiguated. These imperfections prevent machine learning approaches from performing well on automatically generated KGs. Furthermore, many ESE algorithms degrade as the sparsity and unreliability of KGs increases [2, 3]. Advanced ESE methods, especially those that rely on neural networks, are uninterpretable [1]. If a physician can not explain decisions, patients may not trust her and if a journalist can not demonstrate how a certain individual is acting unethically or above the law, a resulting article may lack credibility. Furthermore, uniterpretability may limit the applications of advancements in IR, and more broadly artificial intelligence, as humans "won't trust an A.I. unless it can explain itself."[1]

We introduce Neural Variational Set Expansion (NVSE) to advance the applicability of ESE research. NVSE is an unsupervised model based on Variational Autoencoders (VAEs) that receives a query, uses a Bayesian approach to determine a latent concept that unifies entities in the query, and returns a ranked list of similar entities based on the previously determined unified latent concept. NVSE does not require supervised examples of queries and responses, nor pre-built clusters of entities. Instead, our method only requires sentences with linked entity mentions, i.e. spans of token associated with a KG entity, often included in automatically generated KGs.

[*] Email: `pushpendre@jhu.edu` [†] Work done while the author was at JHU [1] `https://nyti.ms/2hR1S15`

In: Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop on Data Search (DATA:SEARCH18). Co-located with SIGIR 2018, Ann Arbor, Michigan, USA – 12 July 2018, published at http://ceur-ws.org

References

[1] Mitra, B., Craswell, N.: Neural Models for Information Retrieval. ArXiv e-prints (May 2017)

[2] Pujara, J., Augustine, E., Getoor, L.: Sparsity and noise: Where knowledge graph embeddings fall short. In: EMNLP. (2017)

[3] Rastogi, P., Lyzinski, V., Van Durme, B.: Vertex nomination on the cold start knowledge graph. Technical report, Human Language Technology Center of Excellence (2017)