Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short

Jay Pujara and **Eriq Augustine** and **Lise Getoor**
Department of Computer Science
University of California, Santa Cruz
jay@cs.umd.edu, eaugustine@ucsc.edu, getoor@soe.ucsc.edu

Abstract

Knowledge graph (KG) embedding techniques use structured relationships between entities to learn low-dimensional representations of entities and relations. One prominent goal of these approaches is to improve the quality of knowledge graphs by removing errors and adding missing facts. Surprisingly, most embedding techniques have been evaluated on benchmark datasets consisting of dense and reliable subsets of human-curated KGs, which tend to be fairly complete and have few errors. In this paper, we consider the problem of applying embedding techniques to KGs extracted from text, which are often incomplete and contain errors. We compare the sparsity and unreliability of different KGs and perform empirical experiments demonstrating how embedding approaches degrade as sparsity and unreliability increase.

1 Introduction

Recently knowledge graphs (KGs), structured representations of knowledge bases, have become an essential component of systems that perform question-answering (Berant et al., 2013), provide decision support, and enable exploration and discovery (Dong et al., 2014). Initial efforts to create KGs focused on structured information sources or relied extensively on manual curation. However, the diversity of knowledge available on resources like the World Wide Web have spurred many projects that tackle the more difficult task of automatically constructing KGs (Nickel et al., 2016a).

Unfortunately, information extraction approaches for KG construction must overcome complex, unreliable, and incomplete data. Many machine learning methods have been proposed to address the challenge of cleaning and completing KGs. One popular class of methods learn embeddings that translate entities and relationships into a latent subspace, then use this latent representation to derive additional, unobserved facts and score existing facts (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015).

Embedding methods have shown state-of-the-art results on several benchmark datasets. However, by construction, these benchmark datasets differ from data in real KGs. First, benchmark datasets have largely been restricted to the most frequently occurring entities in the KG. However in most KGs, entities are associated with a sparse set of observations. Second, benchmark datasets consist only of highly reliable facts from curated knowledge bases. In contrast, many KG construction projects extract knowledge from noisy data such as text or images, which introduces unreliable information.

In this paper, we evaluate popular KG embedding approaches on KGs that have sparse entities and unreliable candidate facts. We apply embedding methods to an extracted KG and modify existing benchmarks by varying the sparsity and reliability of training data used to learn embedding models. Using this suite of datasets, we characterize where embedding approaches are successful and the conditions that result in degrading results. Based on our insights, we provide recommendations for improving embedding models and identify promising areas of future exploration.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1752–1757
Copenhagen, Denmark, September 7–11, 2017.

Dataset	Triples	$\|E\|$	$\|R\|$	EE	RE	ED	RD	prec
Freebase	1B	124M	15K	14	3.2	16	68K	1
WordNet	380K	116K	27	21	2.3	7	21K	1
NELL1000	92M	4.8M	435	21	4.9	19	210K	0.45
FB15K	592K	15K	1.3K	16	5.1	79	440	1
WN18	151K	40K	18	19	2.1	7	8.4K	1
NELL165	1M	820K	221	25	1.5	3	4.7K	0.35

Table 1: Statistics of knowledge graphs and extracted datasets. Triples are the number of individual facts in the knowledge graph. $\|E\|$ and $\|R\|$ are unique entities and relations in the KG, respectively. EE and RE are measures of entropy, ED and RD measures of density, and prec is the precision of triples.

2 Background and Related Work

Diverse strategies for knowledge base construction include manually-crafted ontologies for common-sense reasoning (Lenat, 1995), community-driven collaborative efforts (Bollacker et al., 2008), ontology-based extraction from structured and textual sources (Mitchell et al., 2015), and "open" approaches that rely on textual information (Mausam et al., 2012). In this paper, we contrast the properties of two knowledge graphs that have clean, human-vetted facts with two knowledge graphs that are extracted from textual data.

Semantically meaningful embeddings of text have been a longstanding topic of study in NLP research (Turney and Pantel, 2010). More recently, knowledge graphs, which capture structured relationships between entities, has inspired methods such as matrix factorization (Riedel et al., 2013), tensor factorization (Nickel et al., 2011), and deep learning (Socher et al., 2013) that embed entities while preserving this relationship structure. We consider four state-of-the-art embedding methods (Bordes et al., 2013; Wang et al., 2014; Nickel et al., 2016b; Nguyen et al., 2016) and assess their performance on knowledge graphs with different properties.

3 Comparing Properties of KGs

In Table 1, we introduce three knowledge graphs and a parallel set of benchmark datasets derived from these KGs. Each KG takes the form of triples that specify a relationship between a subject and an object. The first two KGs, Freebase and WordNet, benefit from human curation that results in precisely defined entities and relationships and highly reliable facts. The third KG, NELL, is extracted from a large Web text corpus using an iterative co-training process and a pre-defined set of relations and types. Due to the iterative nature, NELL is a dynamic dataset and the table reports statistics of the 1000^{th} iteration. FB15K and WN18, derived from Freebase and WordNet, respectively, have been used to train and evaluate many embedding strategies. NELL165, based on an earlier iteration of NELL, has been used as a benchmark for probabilistic models. We compare the vital statistics of these six datasets.

3.1 Size and Sampling

Despite the reliance on curation, Freebase is the largest KG with more facts ($\|T\|$), unique entities ($\|E\|$), and relationship types ($\|R\|$) than others. NELL, is a tenth the size of Freebase with substantially fewer entities and limited relations. WordNet, focused on NLP, is the smallest and expresses only 27 relationships between different words. The derived benchmark datasets are substantially smaller than the source KGs, with the largest, NELL, containing 1M facts. FB15K is generated by sampling a subset of the KG centered around 15K entities. WN18 is generated by restricting to 18 relations. NELL165 performs no sampling, but is limited by the comprehensiveness of patterns learned during training.

3.2 Diversity

To understand the distribution of entities and relationships in the KG, we introduce an entropy-based measure using the probability an entity or relation will occur in a randomly

selected triple. For triples T of the form (s, p, o), relations R, entities E, We define the entity and relation probabilities as the probability that a randomly selected triple will contain a particular relation or entity. More formally, we define these probabilities:

$$P(r) = \frac{|t.p = r|}{\|T\|}; \quad P(e) = \frac{|t.s = e| + |t.o = e|}{\|T\|}$$

Using these definitions, we define:

$$RE = \sum_{r \in R} -P(r) \log P(r)$$

We compute entity entropy (EE) and relation entropy (RE) for each dataset. Higher entropy values indicate more uniform distributions of facts across entities and relations, lower values signal biases in the facts. For example, the low RE values for Freebase and NELL165 are due to an abundance of facts specifying entity types (such as person), relative to other relations between entities. While Freebase has the most facts and entities, these facts are less diverse compared to other KGs. Through sampling, FB15K rebalances Freebase, increasing the diversity of entities and relations. In contrast, WordNet and WN18 have similar diversity statistics. Compared to NELL1000, NELL165 has a more diverse set of entities and a less diverse set of relations. All KGs have much higher EE than RE, since they use a manually defined set of relations but include many diverse entities.

3.3 Sparsity

In addition to diversity, KGs have differing levels of factual information for each entity or relation. One sparsity metric is information density, defined as the average triples per entity or relation. We formally define densities:

$$RD = \frac{\|T\|}{\|R\|}; \quad ED = \frac{2\|T\|}{\|E\|}$$

We compare the datasets using entity density (ED) and relational density (RD). Most datasets have a similar ED, but the benchmark dataset FB15K has much higher entity density while the benchmark dataset NELL165 has a much lower entity density. NELL1000 has the highest RD, since extractions are focused on a small set of relations, while FB15K has a particularly low RD value due to the entity-centric approach to construction. We note that FB15K has much higher ED and much lower RD than parent Freebase, due to the sampling choices made during its construction.

3.4 Reliability

Embedding approaches rely on using facts that are reliable. Human-curated KGs generally have high precision due to strong oversight. In contrast, extracted KGs are far noisier, including erroneous relationships between entities. Extracted KGs are often evaluated on small, manually-labeled evaluation sets to estimate precision. In recent evaluations (Mitchell et al., 2015) using 11K annotations, NELL facts had a precision of ranging from 0.75-0.85 for confident extractions and 0.35-0.45 across the broader set of extractions.

4 Empirical Evaluation

To better understand embedding performance with sparse and unreliable data, we select four popular embedding approaches and perform four empirical analyses. We evaluate embedding techniques TransE (Bordes et al., 2013), TransH (Wang et al., 2014), HolE (Nickel et al., 2016b), and STransE (Nguyen et al., 2016), that use increasingly sophisticated learning methods to represent entities and relations. To learn embeddings, we used the public implementations of Lin et al. (2015); Nickel et al. (2016b); Nguyen et al. (2016). We conduct four experiments to characterize the performance of these embeddings methods. The first set of experiments evaluate the performance of embeddings on the extracted NELL165 KG. The second set of experiments modify the existing FB15K benchmark to isolate the impact of sparsity on embedding quality. The third set of experiments decrease the reliability of FB15K and determine how performance degrades as a result. The final experiments explore the tradeoff between sparsity and reliability by beginning with a sparse trainng set and incrementally adding unreliable triples at differing noise levels.[1]

[1] Code for experiments is available at https://www.github.com/linqs/pujara-emnlp17

Method	AUPRC	F1
Baseline	0.873	0.828
NELL	0.765	0.673
TransH	0.701	0.783
HolE	0.710	0.783
TransE	0.726	0.783
STransE	0.784	0.783
PSL-KGI	**0.891**	**0.848**

Table 2: Embedding performance on the sparse and noisy NELL165 benchmark is poor, failing to beat a baseline that simply selects the top extractions, and substantially underperforming probabilistic models.

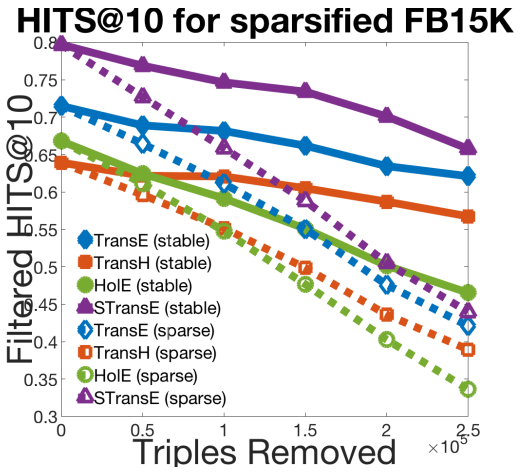

Figure 1: Triples are removed from FB15K to preserve relational density (`stable`, solid) or to increase sparsity (`sparse`, dotted). Sparse training sets have a pronounced impact on the learned embedding, as measured by HITS@10 on the test set.

4.1 Extracted Knowledge Graphs

In Section 3, we noted that the extracted NELL165 dataset is sparse, with fewer (candidate) facts per relation or entity than the FB15K benchmarks. Moreover, the precision of these candidates can be far lower than benchmark datasets. To evaluate whether embeddings can succeed under such challenging conditions, we applied four state-of-the-art embedding techniques,

We evaluated all methods on 4.5K manually-labeled facts (Jiang et al., 2012), reporting the area under the precision-recall curve (AUPRC) and the F1 score, computed with parameters that maximize performance on the labeled training set. We compare against a baseline that simply applies a threshold to NELL extractor confidences (but cannot score novel facts), the NELL promotion strategy, and a probabilistic approach PSL-KGI (Pujara et al., 2015), that reasons collectively about KG facts using ontological constraints and supports open-world reasoning. The results, in Table 2, suggest that embedding approaches cannot cope with the sparse and low-quality extractions, performing more poorly than the baseline approaches and substantially trailing the probabilistic model. In the next two experiments, we analyze whether this failure can be attributed to sparsity or sensitivity to noise.

4.2 Sensitivity to Sparsity

One potential explanation for the lackluster performance of embedding approaches on extracted KGs is the sparsity of these datasets. To assess the impact of sparsity on the quality of learned embeddings, we remove triples from FB15K using two different techniques. The first technique, `sparse`, removes triples uniformly at random, with a constraint that such removal does not eliminate any entity or relation from the dataset. The second technique, `stable`, removes all triples for a particular relation, leaving other relations intact. `stable` is calibrated so that the training set size does not vary more than 2% between techniques.

Fig. 1 shows the filtered hits@10 metric (proportion of correct triples in top ten triples excluding training data) for both `sparse` and `stable` using the TransE, TransH, HolE, and STransE embeddings. Performance universally decreases as the training set diminishes. However, in the `sparse` treatment, performance deteriorates much more rapidly than in `stable`. Our experiments show that more complex representations such as TransH and HolE suffer more from sparsity, while TransE and the more sophisticated STransE have somewhat better performance. Ultimately, when half the triples have been randomly removed, corresponding to a (relatively high) RD value of 220, the `stable` outperforms `sparse` by as much as 60%. The contrast between a dense set of facts for each relation (`stable`) and a sparse set of relational training data is a vivid demonstration that embedding quality relies on dense training data.

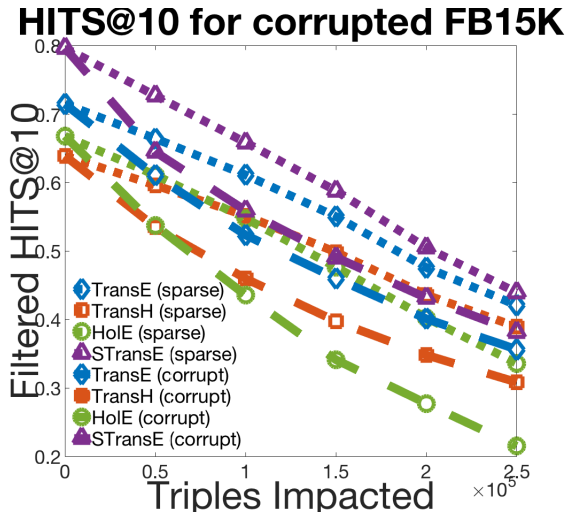

Figure 2: Randomly corrupting triples (`corrupt`, dashed) during training decreases embedding quality relative to randomly removing triples (`sparse`, dotted).

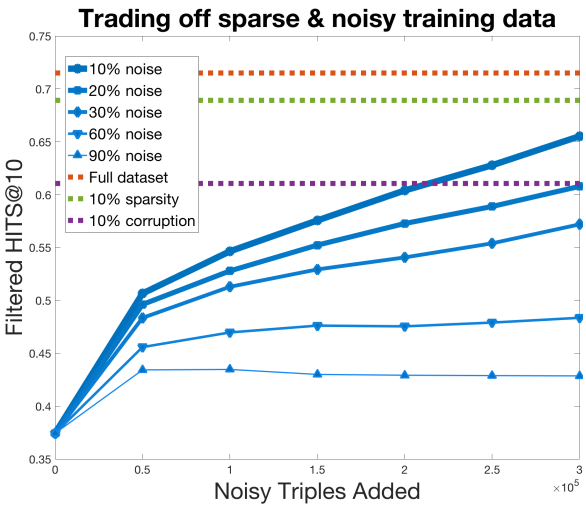

Figure 3: Starting with a sparse training set, adding unreliable triples can help embedding performance recover if the noise level is low.

4.3 Sensitivity to Unreliability

Beyond sparsity, candidate facts generated by knowledge extraction approaches can also be unreliable. To understand the sensitivity of embedding techniques to noise, we modified the FB15K dataset to include unreliable triples. Our approach to introducing noise, `corrupt` involved "corrupting" triples, substituting a replacement entity or relation for the true subject, predicate or object. The embedding approach is then trained with a corrupted version of the benchmark. Fig. 2 show how the Hits@10 metric suffers as increasing numbers of facts are either corrupted (`corrupt`) or removed (`sparse`). We find that across all methods, removing training data is better than providing incorrect training data to the learning algorithm, but surprisingly the deficit between `sparse` and `corrupt` remains relatively stable across all embeddings.

4.4 Trading off Sparsity and Noise

In many real-world scenarios, constructing a KG requires navigating a tradeoff between sparsity and noise. A sparse, high-quality set of extractions may be insufficient to learn meaningful embeddings. However, the benefit of incorporating additional, unreliable facts may also be questionable. We explore this tradeoff by randomly removing $300K$ triples from FB15K and incrementally adding unreliable triples at differing noise levels, where noise measures the probability a newly-added training triple is corrupted. We generate training sets for each noise level and size, train TransE, and compute the filtered Hits@10 metric on the test set. Fig. 2 shows all embeddings have an initial benefit from new training data, but noise level dictates the improvement as more data is introduced. For low noise settings, performance climbs steadily, while higher noise results in plateauing or diminishing performance. Surprisingly, even with 90% noise embeddings demonstrate a small net improvement, suggesting that for embedding methods a large, unreliable corpus may be better than an extremely sparse, high-quality one.

5 Conclusion

In this paper, we analyze several knowledge graphs and discuss key metrics for diversity, sparsity, and unreliability in realistic KGs. Our experimental evaluation concludes that KG embeddings are sensitive to sparse and unreliable data, and perform poorly on KGs extracted from text. These findings suggest a rich area of future research, determining new strategies to extend embeddings to cope with sparse and unreliable data. Three promising approaches include revising the closed-world assumption frequently used in training embeddings, combining embeddings and collective probabilistic models that perform well on extracted KGs, and devising an optimization approach for embeddings that exploits confidence from knowledge extraction systems.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *SIGMOD*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM.

Shangpu Jiang, Daniel Lowd, and Dejing Dou. 2012. Learning to Refine an Automatically Extracted Knowledge Base Using Markov Logic. In *ICDM*.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *AAAI*.

Mausam, Michael D. Schmitz, Robert E. Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP*.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-Ending Learning. In *AAAI*.

Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. 2016. STransE: A Novel Embedding Model of Entities and Relationships in Knowledge Bases. In *NAACL-HLT*.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016a. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1).

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016b. Holographic Embeddings of Knowledge Graphs. In *AAAI*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *ICML*.

Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2015. Using Semantics & Statistics to Turn Data into Knowledge. *AI Magazine*, 36(1):65–74.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *NAACL-HLT*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *JAIR*, 37(1).

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *AAAI*.