Row-less Universal Schema

Patrick Verga & Andrew McCallum
College of Information and Computer Sciences
University of Massachusetts Amherst
{pat, mccallum}@cs.umass.edu

Abstract

Universal schema jointly embeds knowledge bases and textual patterns to reason about entities and relations for automatic knowledge base construction and information extraction. In the past, entity pairs and relations were represented as learned vectors with compatibility determined by a scoring function, limiting generalization to unseen text patterns and entities. Recently, 'column-less' versions of Universal Schema have used compositional pattern encoders to generalize to all text patterns. In this work we take the next step and propose a 'row-less' model of universal schema, removing explicit entity pair representations. Instead of learning vector representations for each entity pair in our training set, we treat an entity pair as a function of its relation types. In experimental results on the FB15k-237 benchmark we demonstrate that we can match the performance of a comparable model with explicit entity pair representations using a model of attention over relation types. We further demonstrate that the model performs with nearly the same accuracy on entity pairs never seen during training.

1 Introduction

Automatic knowledge base construction (AKBC) is the task of building a structured knowledge base (KB) of facts using raw text evidence, and often an initial seed KB to be augmented (Carlson et al., 2010; Suchanek et al., 2007; Bollacker et al., 2008). Extracted facts about entities and their relations are useful for many down-stream tasks such as question answering and query understanding. An effective approach to AKBC is Universal Schema, in which relation extraction is modeled as a matrix factorization problem wherein each row of the matrix is an entity pair and each column represents a relation between entities. Relations derived from a KB schema and from free text are thus embedded into a shared space allowing for a rich representation of KB relations, the union of all KB schemata.

This formulation is still limited in terms of its generalization, however. In its original form, Universal Schema can reason only about entity pairs and text relations explicitly seen at train time; it cannot predict relations between new entity pairs. In this work we present a 'row-less' extension of Universal Schema. Rather than representing each entity pair with an explicit dense vector, we encode entity pairs as aggregate functions over their relation types. This allows Universal Schema to form predictions for all entity pairs regardless of whether that pair was seen during training, and provides a direct connection between the prediction and its provenance.

Many models exist which address this issue by operating at the level of entities rather than entity pairs. A knowledge base is naturally described as a graph, in which entities are nodes and relations are labeled edges (Suchanek et al., 2007; Bollacker et al., 2008). In the case of *knowledge graph completion*, the task is akin to link prediction, assuming an initial set of (*s, r, o*) triples. See Nickel et al. (2015) for a review. No accompanying text data is necessary, since links can be predicted using properties of the graph, such as transitivity. In order to generalize well, prediction is often posed as low-rank matrix or tensor factorization. A variety of model variants have been suggested, where the probability of a given edge existing depends on a multi-linear form (Nickel et al., 2011; García-Durán et al., 2015; Yang et al., 2015; Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015), or non-linear interactions between s, r, and o (Socher et al., 2013). These entity-based models have recall advantages over entity pairs. The model can predict relations between any two entity pairs in the absence of other information such as the pair's contextual occurrence in text.

However, entity models have been shown to be less precise than entity pair models when text is used to augment knowledge base facts. Both Toutanova et al. (2015) and Riedel et al. (2013) observe that the entity pair model outperforms entity models in cases where the entity pair was seen at training time. Since Universal

Schema leverages large amounts of unlabeled text we desire the benefits of entity pair modeling, and row-less Universal Schema facilitates learning entity pairs without the drawbacks of the traditional one-embedding-per-pair approach.

In this paper we investigate Universal Schema models without explicit entity pair representations. Instead, entity pairs are represented using an aggregation function over their relation types. This allows our model to naturally make predictions about any entity pair in new textual mentions, regardless of whether they were seen at train time additionally giving the model direct access to provenance. We show that an attention-based aggregation function outperforms several simpler functions and matches a model using explicit entity pairs. We then demonstrate that these 'row-less' models accurately predict on entity pairs unseen during training.

2 Background

2.1 Universal Schema

The Universal Schema (Riedel et al., 2013) approach to AKBC jointly embeds any number of KB and text corpora into a shared space to jointly reason over entities and their relations (Figure 1). The problem of relation extraction is posed as a matrix completion task where rows are entity pairs and columns are KB relations and textual patterns. The matrix is decomposed into two low-rank matrices resulting in embeddings for each entity pair, relation, and textual pattern. Reasoning is then performed directly on these embeddings.

Riedel et al. (2013) proposed several model variants operating on entities and entity pairs, and subsequently many other extensions have been proposed (Yao et al., 2013; Gardner et al., 2014; Neelakantan et al., 2015; Rocktaschel et al., 2015). Recently, Universal Schema has been extended to encode compositional representations of textual relations (Toutanova et al., 2015; Verga et al., 2016) allowing it to generalize to all textual patterns and reason over arbitrary text.

2.2 'Column-less' Universal Schema

The original Universal Schema approach has two main drawbacks: similar patterns do not share statistics, and the model is unable to make predictions about textual patterns not explicitly seen at train time.

Recently, 'column-less' versions of Universal Schema have been proposed to address these issues (Toutanova et al., 2015; Verga et al., 2016). These models learn compositional pattern encoders to parameterize the column matrix in place of directly embedding textual patterns. Compositional Universal Schema facilitates more compact sharing of statistics by composing similar patterns from the same sequence of word embeddings – the text
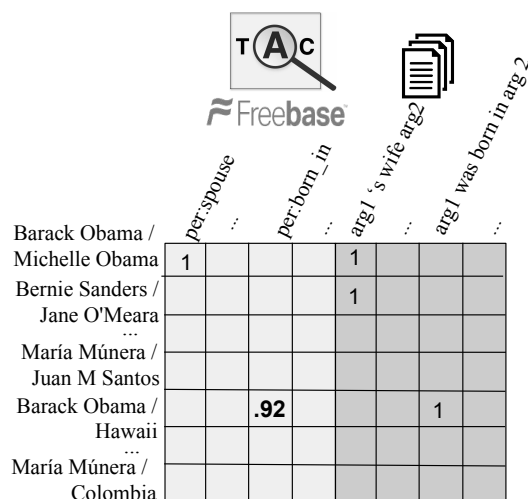

Figure 1: Universal schema represents relation types and entity pairs as a matrix. 1s denote observed training examples and the bolded .92 is predicted by the model.

patterns 'lives in the city' and 'lives in the city of' no longer exist as distinct atomic units. More importantly, Compositional Universal Schema can thus generalize to all possible textual patterns, facilitating reasoning over arbitrary text at test time.

3 Model

3.1 'Row-less' Universal Schema

While column-less Universal Schema addresses reasoning over arbitrary textual patterns, it is still limited to reasoning over entity pairs seen at training time. Verga et al. (2016) approach this problem by using Universal Schema as a sentence classifier – directly comparing a textual relation to a KB relation to perform relation extraction. However, this approach is unsatisfactory for two reasons. The first is that this creates an inconsistency between training and testing, as the model is trained to predict compatibility between entity pairs and relations and not relations directly. Second, it considers only a single piece of evidence in making its prediction.

We address both of these concerns in our 'row-less' Universal Schema. Rather than encoding each entity pair explicitly, we take the compositional approach of encoding entity pairs as an aggregation over their observed relation types (Figure 2). A learned entity pair embedding can be seen as a summarization of all relation types for which that entity pair was seen. Rather than learn this summarization as a single embedding, we reconstruct an entity pair representation from an aggregate of its relation types, essentially learning a mixture model rather than a single centroid.
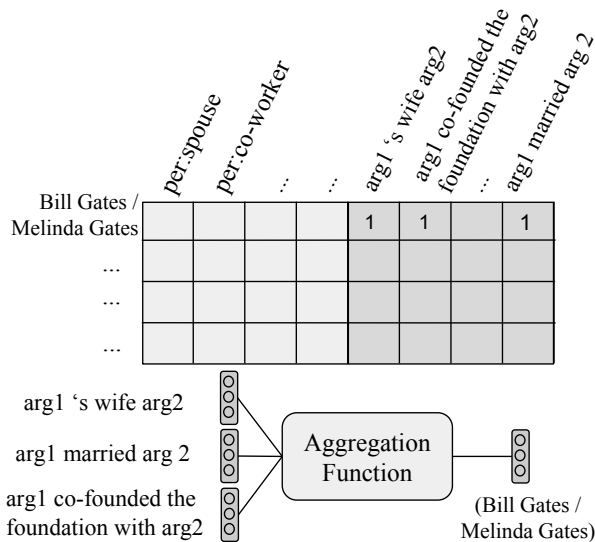
Figure 2: Row-less Universal Schema encodes an entity pair as an aggregation of its observed relation types.

3.2 Aggregation Functions

In this work we examine several aggregation functions. **Mean Pool** creates a single centroid for the entity pair by averaging all of its relation vectors. While this intuitively makes sense as an approximation for the explicit entity pair representation, averaging large numbers of embeddings can lead to a noisy signal. **Max Pool** also creates a single centroid for the entity pair by taking a dimension-wise max over the observed relation type vectors. Both mean pool and max pool are query-independent and form the same representation for the entity pair regardless of the query relation.

We also examine two query-specific aggregation functions. These models are more expressive than a single vector that is forced to to act as a centroid to all possible relation types an entity pair can take on. For example, the entity pair Bill and Melinda Gates could hold the relation 'per:spouse' or 'per:co-worker'. A query-specific aggregation mechanism can produce separate representations for this entity pair dependent on the query.

The **Max Relation** aggregation function represents the entity pair as its most similar relation to the query vector of interest. This model has the advantage of creating a query-specific entity pair representation but is more susceptible to noisy training data as a single incorrect piece of evidence could be used to form a prediction.

Finally, we look at an **Attention** aggregation function over relation types (Figure 3) which is similar to a single-layer memory network (Sukhbaatar et al., 2015) . In this model the query is scored with an input representation of each relation type followed by a softmax, giving a

weighting over each relation type. This output is then used to get a weighted sum over a set of output representations for each relation type resulting in a query-specific vector representation of the entity pair. The model pools relevant information over the entire set of relations and selects the most salient aspects to the query relation.

3.3 Training

Riedel et al. (2013) use Bayesian Personalized Ranking (BPR) (Rendle et al., 2009) to train their Universal Schema models. BPR ranks the probability of observed triples above unobserved triples rather than explicitly modeling unobserved edges as negative. Each training example is an entity pair/relation type triple observed in the training text corpora or KB. Rather than BPR, Toutanova et al. (2015) use a sampled softmax criterion where they use 200 negative samples [1] to approximate the negative log likelihood. Results shown were obtained using sampled softmax which outperformed BPR in our early experiments.

Our training procedure for relation-only Universal Schema model is similar to the original model. We first pool all of the observed triples in our training data creating entity pair-specific relation sets R^{Ep}. We remove entity pairs with only a single observed relation type. We then construct training examples for each observed relation type of an entity and an aggregation of all other relation types observed with that entity; for each observed relation type $r_i \in R^{Ep}$, we construct a positive training example $(r_i, \{R^{Ep} \setminus r_i\})$. We randomly sample a different relation to act as the negative sample.

All models were implemented in Torch[2] and were trained using Adam (Kingma and Ba, 2015). They each were trained with embedding dimension 25 and used 200 negative samples except for max pool which performed better with two negative samples. The entity pair model used a batch size 1024, $\ell_2 = 1e\text{-}8$, $\epsilon = 1e\text{-}4$, and learning rate .01. The aggregation models all used batch size 4096, $\ell_2 = 0$, $\epsilon = 1e\text{-}8$, and learning rate .01. The column vectors were initialized with the columns learned by the entity pair model. Randomly initializing the query encoders and tying the output and attention encoders performed better and all results use this method. Models were tuned to maximize mean reciprocal rank (MRR) on the validation set with early stopping.

---

[1] Many past papers restrict negative samples to be of the same type as the positive example. We simply sample uniformly from the entire set of entities that would form a valid entity pair.

[2] https://github.com/patverga/torch-relation-extraction
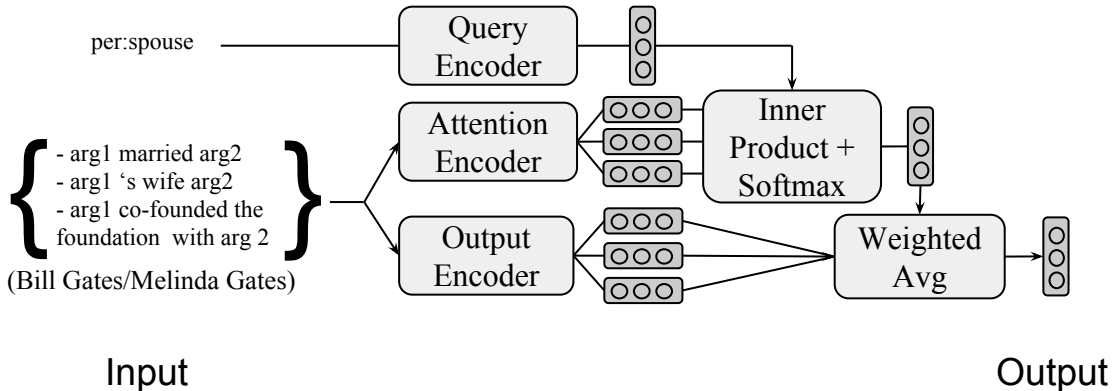
Input — **Output**

Figure 3: In the Attention model the query is dotted with an input representation of each relation type followed by a softmax, giving a weighting over each relation type. This output is then used to get a weighted sum over a set of output representations for each relation type. The result is a query-specific vector representation of the entity pair. The Max Relation model simply takes the max dot product rather than a softmax and weighted average.

4 Experimental Results

4.1 Data and Evaluation

We evaluate our models on the FB15k-237 dataset from Toutanova et al. (2015). The data is composed of a small set of 237 Freebase relations and approximately 4 million textual patterns from Clueweb with entities linked to Freebase (Gabrilovich et al., 2013). In past studies, for each (subject, relation, object) test triple, negative examples are generated by replacing the object with all other entities, filtering out triples that are positive in the data set. The positive triple is then be ranked among the negatives. In our experiments we limit the possible generated negatives to those entity pairs that have textual mentions in our training set. This way we can evaluate how well the model classifies textual mentions as Freebase relations. We also filter textual patterns with length greater than 35. We report the percentage of positive triples ranked in the top 10 amongst their negatives as well as the MRR scaled by 100.

4.2 Results

Our results are shown in Table 1. The model with explicit entity pair representations outperforms the 'row-less' variant by 2.9 MRR and 2.7% Hits@10. However, these results do show that relation aggregation is competitive with entity pair embeddings and it is possible to have a Universal Schema model without entity embeddings. With further experimentation we are confident that the 'row-less' model will perform on-par or better than the entity pair model.

The max model performs competitively with the attention model. This is not entirely surprising as it is a simpli-

fied version of the attention model. Further, the attention model reduces to the max relation model for entity pairs with only a single observed relation type. In our data, 64.8% of entity pairs have only a single observed relation type and 80.9% have 1 or 2 observed relation types.

We also explore the models' abilities to predict on unseen entity pairs (Table 2). We remove all training examples that contain a positive entity pair in either our validation or test set. We use the same validation and test set as in Table 1. The entity pair model predicts random relations as it is unable to make predictions on unseen entity pairs. The max pool and mean pool each suffer approximately 30% relative decrease in MRR and 20% decrease in Hits@10.

Both the max relation and attention models perform nearly as well whether the entity pairs were observed during training or not. The attention model performs slightly better than the max relation model in this scenario.

Model	MRR	Hits@10
Entity-pair Embeddings	31.85	51.72
Mean Pool	25.89	45.94
Max Pool	29.61	49.93
Max Relation	31.71	**51.94**
Attention	**31.92**	51.67

Table 1: The percentage of positive triples ranked in the top 10 amongst their negatives as well as the mean reciprocal rank (MRR) scaled by 100 on a subset of the FB15K-237 dataset. Negative examples were restricted to entity pairs that occurred in the KB or text portion of the training set.

Model	MRR	%dec	Hits@10	%dec
Entity-pair Embeddings	5.23	83.6	11.94	76.9
Mean Pool	18.10	30.1	35.76	22.2
Max Pool	20.80	29.8	40.25	19.4
Max Relation	28.46	10.3	48.15	7.3
Attention	**29.75**	**6.8**	**49.69**	**3.8**

Table 2: Predicting entity pairs that were not seen at train time. The percentage of positive triples ranked in the top 10 amongst their negatives as well as the mean reciprocal rank (MRR) scaled by 100 on a subset of the FB15K-237 dataset. Also shown is the percent relative decrease in MRR and Hits@10 between Table 1 (entity pairs seen during training) and this table (entity pairs unseen during training).

5 Conclusion

In this paper we explore a row-less extension of Universal Schema that forgoes explicit entity pair representations for an aggregation function over relation types. This extension allows prediction between all entity pairs in new textual mentions – whether seen at train time or not – and also provides a natural connection to the provenance supporting the prediction.

In this work we show that an aggregation function based on query-specific attention over relation types outperforms query independent aggregations. We show that aggregation models are able to predict on par with entity pair models for seen entity pairs and, in the case of attention, suffer very little loss for unseen entity pairs.

We also limited our pattern encoders to lookup-tables. In future work we will combine the column-less and row-less approach to make a fully compositional Universal Schema model. This will allow Universal Schema to generalize to all new textual patterns and entity pairs.

References

[Bollacker et al.2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

[Bordes et al.2013] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*.

[Carlson et al.2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and A. 2010. Toward an architecture for never-ending language learning. In *In AAAI*.

[Gabrilovich et al.2013] Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by*, 5.

[García-Durán et al.2015] Alberto García-Durán, Antoine Bordes, Nicolas Usunier, and Yves Grandvalet. 2015. Combining two and three-way embeddings models for link prediction in knowledge bases. *CoRR*, abs/1506.00999.

[Gardner et al.2014] Matt Gardner, Partha Talukdar, Jayant Krishnamurthy, and Tom Mitchell. 2014. Incorporating vector space similarity in random walk inference over knowledge bases. In *Empirical Methods in Natural Language Processing*.

[Kingma and Ba2015] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations (ICLR)*.

[Lin et al.2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI*.

[Neelakantan et al.2015] Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015. Compositional vector space models for knowledge base completion. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

[Nickel et al.2011] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*.

[Nickel et al.2015] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction. *arXiv preprint arXiv:1503.00759*.

[Rendle et al.2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press.

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

[Rocktaschel et al.2015] Tim Rocktaschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[Socher et al.2013] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*.

[Suchanek et al.2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*.

[Sukhbaatar et al.2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439.

[Toutanova et al.2015] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[Verga et al.2016] Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[Wang et al.2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1112–1119. Citeseer.

[Yang et al.2015] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *International Conference on Learning Representations 2014*.

[Yao et al.2013] Limin Yao, Sebastian Riedel, and Andrew McCallum. 2013. Universal schema for entity type prediction. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 79–84. ACM.