知識グラフ埋め込みの PU 学習

垣渕 太成 [*1]　林 克彦 [†2]　駒谷 和範 [‡1]

[1] 大阪大学 産業科学研究所
[2] 群馬大学 社会情報学部

1 はじめに

知識グラフは，事実（Fact）を集積した巨大なデータセットであり，Fact は，Entity e とその間の Relation r の三つ組（Triple）形式 (e_i, r_k, e_j) で表現される．ここで前者の Entity は Subject，後者は Object と呼ばれる．知識グラフの多くは Fact の欠損を含み，これを自動予測する知識グラフ補完は実用上重要なタスクである．そして，知識グラフ埋め込みはこのタスクに対する有効なアプローチの一つと考えられている．埋め込みの学習は Logistic 損失や Cross Entropy（CE）損失に基づいて，Fact とそれ以外の Triple を分類するように行うのが一般的である．

知識グラフでは Fact 以外の Triple の正負が不明であるため，知識グラフ埋め込みの学習を行うには負例をどう定義するかが重要となる．既存の学習法では，Fact の Subject あるいは Object を別の Entity に入れ替えた Triple を負例として定義する方法がよく用いられている．しかし，この負例の定義が正確である保証はなく，潜在的な Fact を負例として扱ってしまう危険が常に伴う．また，Fact の近傍のみを学習対象とするため，局所的で偏った学習を行う問題も指摘できる．分野ではモデルの性能を Fact 近傍のランキングで評価することが主流であるため，これらの問題は顕在化してこなかったが，本来，知識グラフ補完の性能を評価するには与えられた Triple が Fact か否かを判定する分類評価が適切である．

そこで，本稿ではまず，従来の学習法によるモデルの分類性能を再調査する．図 1 は Kinship データセットに対する既存モデルの性能を PR（Precision and Recall）曲線で示しており，この結果は，Precision が早期に下がって，分類性能が極めて低いことを示している．これは負例が正例と判定される割合が高いことに起因するもので，この結果から，従来の学習法における負例の定義を根本的に見直す必要性があると結論付けられる．

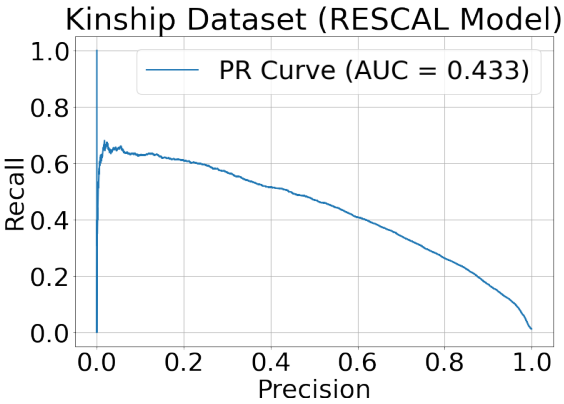

図 1　Kinship データセットの Fact から 3 割を除いて，CE 損失による既存の埋め込み学習（RESCAL モデルを使用）を行い，除いた Fact に対して PR 曲線を計算した．

よって，本稿では，**PU（Positive and Unlabeled）学習**の知見を用いた知識グラフ埋め込みの新たな学習法を提案する．PU 学習では，正例以外全て Unlabeled であるデータセットに一部正例が含まれていると仮定した上で正と負への二値分類を行うため，明示的な負例を定義する必要がない．本稿では知識グラフ埋め込みの代表的なモデルである RESCAL を用い，Alternating Least Square（ALS）による最適化法を定式化した．これを既存手法と比較したところ，分類精度の大きな向上が見られた．

2 関連研究

2.1 知識グラフ埋め込みの学習

初期には二乗誤差損失を用いて ALS で最適化する手法 [1] が提案されたが，これには Unlabeled なデータを全て負例として扱うという欠点がある．現在，事前知識 [2] や敵対学習 [3] によって Fact 近傍で負例の可能性が高い Triple をサンプリングし，Logistic 損失や CE 損失に基づいて学習を行う方法が一般的となっている．しかし，任意の知識グラフ

* kakibuchi@ei.sanken.osaka-u.ac.jp
† khayashi0201@gmail.com
‡ komatani@sanken.osaka-u.ac.jp

に対して正確な負例を定義する方法は自明でない．

2.2 PU 学習

PU 学習 [4] は，正例と Unlabeled で構成されるデータセットに対して，Unlabeled に一部正例が含まれている状況を仮定し，正負の二値分類を行うもので，土地被覆分類 [5] など，負例の抽出が難しい状況で利用されている．文献 [6] は正負のラベルの誤りを考慮した上での学習に取り組んでいる．正例に誤りはないが，負例のデータに正クラスが誤って混在している状況は PU 学習の仮定と等しい．文献 [7] は，行列分解における PU 学習に取り組んでおり，本研究はそのテンソル分解への拡張とも言える．

3 知識グラフ埋め込みの PU 学習

3.1 記法

ベクトル，行列，テンソルをそれぞれ小文字の太文字 $\boldsymbol{a}$，大文字の太文字 $\boldsymbol{A}$，カリグラフィック体 $\mathcal{A}$ で表す．行列 $\boldsymbol{A}$ の i 番目の行ベクトルを $\boldsymbol{a}_i^T$，行列の各要素を a_{ij}，テンソルの各要素を x_{ijk} で表す．$\mathcal{X}_{[i::]}, \mathcal{X}_{[:j:]}, \mathcal{X}_{[::k]}$ はそれぞれテンソル 1, 2, 3 次元の i, j, k 番目のスライスを表す．また $vec(\boldsymbol{A})$ は行列 $\boldsymbol{A}$ のベクトル表示 $(a_{11}, ..., a_{1n}, ..., a_{m1}, ..., a_{mn})^T$ である．行列のフロベニウスノルムを $\|\cdot\|_F$，$\otimes$ はクロネッカー積，$\times_n$ はテンソルの n モード積を表す．

3.2 従来法

知識グラフ埋め込みでは Fact の分布を近似する潜在的な特徴表現の獲得を目指す．以下では，Fact の集合を $\mathcal{F}$，Entity の集合を $\mathcal{E}$，Relation の集合を $\mathcal{R}$ で表す．このとき，知識グラフは二値のテンソル $\mathcal{X} \in \{0,1\}^{|\mathcal{E}|\times|\mathcal{R}|\times|\mathcal{E}|}$ の形で表現でき，各要素 x_{ijk} は $(e_i, r_k, e_j) \in \mathcal{F}$ なら 1，そうでなければ 0 をとる．

RESCAL モデルでは Entity e_i を d 次元の縦ベクトル $\boldsymbol{e}_i$，Relation r_k を $d \times d$ 次元の行列 $\boldsymbol{R}_k \in \mathbb{R}^{d\times d}$ で表現し，隣接行列 $\mathcal{X}_{[::k]}$ を以下のように近似する．

$$\mathcal{X}_{[::k]} \approx \boldsymbol{E}\boldsymbol{R}_k\boldsymbol{E}^T \quad (k = 1, 2, ..., |\mathcal{R}|).$$

Triple に対するスコア関数は以下となる．

$$\theta_{ijk} = \boldsymbol{e}_i^T \boldsymbol{R}_k \boldsymbol{e}_j.$$

他にも様々なモデルが提案されているが，RESCAL は表現力が完全であり [8]，現在でも標準的なモデルとして使用されている [9]．

一般的な知識グラフ埋め込みでは，スコア関数に対する損失を $\ell(\theta_{ijk}, x_{ijk})$ として目的関数は以下のように定義される．

$$L = \sum_{(e_i,r_k,e_j)\in\mathcal{F}} \ell(\theta_{ijk}, 1) + \sum_{(e_i,r_k,e_j)\in\mathcal{D}^-} \ell(\theta_{ijk}, 0) \qquad (1)$$

where

$$\mathcal{D}^- = \{(e_{i'}, r_k, e_j) | e_{i'} \in \mathcal{E} \wedge e_{i'} \neq e_i \wedge (e_i, r_k, e_j) \in \mathcal{F}\}$$
$$\cup \{(e_i, r_k, e_{j'}) | e_{j'} \in \mathcal{E} \wedge e_{j'} \neq e_j \wedge (e_i, r_k, e_j) \in \mathcal{F}\}.$$

$\mathcal{D}^-$ は学習で負に分類される Triple の集合である．これは正例以外の定義可能な Triple 全てを含むわけではなく，Fact(e_i, r_k, e_j) に対して，$(e_i, r_k, ?), (?, r_k, e_j)$ となるものだけが含まれる．

3.3 知識グラフ埋め込みの PU 定式化

PU 学習 [6, 7] の知見を用いて，知識グラフ埋め込みの新たな学習法を定式化する．知識グラフに本来存在する Fact 数から ρ の割合でランダムに欠損が起こると仮定する．すなわち ρ は，真の Fact の集合を $\hat{\mathcal{F}}$，現在の集積されている Fact の集合を $\mathcal{F}$ として $\rho = 1 - |\mathcal{F}|/|\hat{\mathcal{F}}|$ と定義される．$\hat{\mathcal{F}}$ に基づいて損失関数 ℓ で分類した場合の損失を近似するような損失関数 $\tilde{\ell}$ を定義し，目的関数は以下のように修正できる．

$$L = \sum_{(e_i,r_k,e_j)\in\mathcal{F}} \tilde{\ell}(\theta_{ijk}, 1) + \sum_{(e_i,r_k,e_j)\notin\mathcal{F}} \tilde{\ell}(\theta_{ijk}, 0) \qquad (2)$$

$$\text{where} \begin{cases} \tilde{\ell}(\theta_{ijk}, 1) = \frac{\ell(\theta_{ijk},1) - \rho\ell(\theta_{ijk},0)}{1-\rho} \\ \tilde{\ell}(\theta_{ijk}, 0) = \ell(\theta_{ijk}, 0) \end{cases}$$

知識グラフ $\hat{\mathcal{F}}$ に対応するテンソルが $\hat{\mathcal{X}}$，その要素を $\hat{x}_{ijk}$ とすると，損失について以下の等式が成り立つ．

$$\mathbb{E}\left[\tilde{\ell}(\theta_{ijk}, x_{ijk})\right] = \mathbb{E}\left[\ell(\theta_{ijk}, \hat{x}_{ijk})\right]$$

知識グラフにおいて ρ の値は不明だが，ハイパーパラメータとして与えた ρ が実際に Fact の欠損している割合と等しい時，欠損を含まない知識グラフを学習した場合の損失に近似できる．

3.4 ALS による最適化

本研究では，損失 ℓ に二乗誤差 $(x_{ijk} - \theta_{ijk})^2$ を採用し，ALS による最適化を行う．このとき，

$$\tilde{\ell}(\theta_{ijk}, 1) = \left(\theta_{ijk} - \frac{1}{1-\rho}\right)^2 - \frac{\rho}{(1-\rho)^2}$$

と書ける．第二項目は定数項であるから，

$$\text{argmin } L = \text{argmin} \left[\sum_{x_{ijk}=1} \left(\theta_{ijk} - \frac{1}{1-\rho}\right)^2 + \sum_{x_{ijk}=0} \theta_{ijk}^2\right]$$

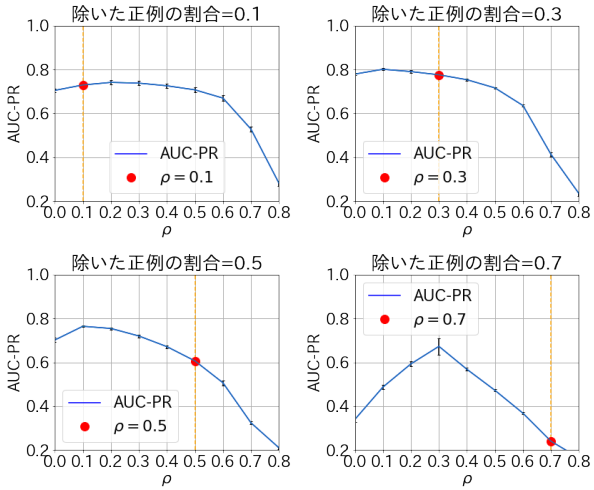

図 2　AUC-PR と ρ の関係（Kinship）.

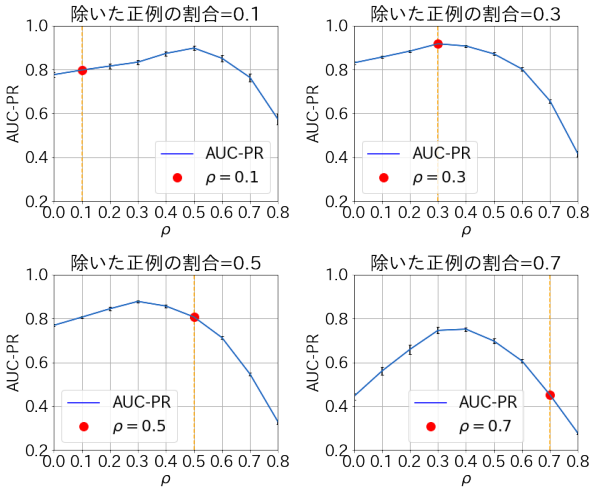

図 3　AUC-PR と ρ の関係（UMLS）.

が成立する．よって，要素が 0 と $1/(1-\rho)$ のテンソル分解の問題に帰着する．以下では，この二値のテンソル $\mathcal{X}'$ を $\mathcal{X}' := 1/(1-\rho)\mathcal{X}$ と定義する．

ALS は行列・テンソル分解において使用されるアルゴリズムであり，分解する行列の片方を固定した上でもう一方の行列をコスト関数の偏微分が 0 になるよう更新し，これを交互に繰り返す．

本研究では RESCAL モデルを使用した．正則化項を加えたコスト関数と ALS 更新式を以下に示す．

$$L = \frac{1}{2}\sum_k \|\mathcal{X}'_{[::k]} - \boldsymbol{E}\boldsymbol{R}_k\boldsymbol{E}^T\|_F^2 + \lambda\left(\|\boldsymbol{E}\|_F^2 + \sum_k \|\boldsymbol{R}_k\|_F^2\right)$$

更新式

$$\boldsymbol{E} \leftarrow \left[\sum_k \mathcal{X}'_{[::k]}\boldsymbol{E}\boldsymbol{R}_k + \mathcal{X}'^T_{[::k]}\boldsymbol{E}\boldsymbol{R}_k^T\right]\left[\sum_k \boldsymbol{B}_k + \boldsymbol{C}_k + \lambda\boldsymbol{I}\right]^{-1}$$

where $\boldsymbol{B}_k = \boldsymbol{R}_k\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{R}_k^T$, $\boldsymbol{C}_k = \boldsymbol{R}_k^T\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{R}_k$

$$\boldsymbol{R}_k \leftarrow (\boldsymbol{Z}^T\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}vec(\mathcal{X}'_{[::k]})$$

where $\boldsymbol{Z} = \boldsymbol{E} \otimes \boldsymbol{E}$.

ALS を用いる利点として，大規模な知識グラフへの対応が挙げられる．文献 [10] では，知識グラフのスパース性を利用し 300 万の Entity，40 の Relation，および 7000 万の Fact で構成される知識グラフの埋め込みに成功している．

4　実験

4.1　分類精度の評価

3 節で提案した PU 学習に基づく定式化を実験的に検証する．使用したデータセットは Kinship と UMLS で，共に Fact が全て集積されておりそれ以外の Triple は全て負例である性質を持つ [11].

データセットから Fact をランダムに一定割合 $\{0.1, 0.3, 0.5, 0.7\}$ 除き残りを学習に与え，それ以外の全 Triple に対し分類精度をテストする．検証の目的は以下の三つである．

1. PU 学習により分類精度は向上するか.
2. ρ と除いた正例の割合の関係．両者が等しい時に分類精度が最も高くなるか.
3. 従来法との比較.

目的 1, 2 の検証のため，ハイパーパラメータ ρ ごとに AUC-PR の値を調べた（図 2, 3）．ここで，次元数 100，正則化項 $\lambda = 0.0001$ で固定し 10 分割交差検証を行った．Fact を除いた割合によらず，$\rho > 0.0$ の適当な値で $\rho = 0.0$（PU なし）と比較して分類精度が向上している．一方，理論的には除いた正例の割合に対し ρ が等しい場合（図赤点）で AUC が最も高くなることが予想されるが，そのような傾向は得られなかった．一般的な分類問題と異なり知識グラフ埋め込みでは潜在的な特徴表現を通じて各 Triple のスコアが相互に影響を及ぼすため，理論通りの学習が行われてない可能性があり，知識グラフ埋め込みにより適した PU 学習の定式化は今後の課題である．

続いて従来法と提案手法を比較する．従来法には RESCAL モデル，目的関数に式 (1)，損失関数に Sigmoid Cross Entoropy を用いて SGD（Stochastic Gradient Descent）で最適化を行った．ハイパーパラメータは次元数 $d \in \{30, 50, 100\}$，学習率は $\eta \in \{0.005, 0.003, 0.001\}$ から Grid Search を行った．Epoch は 50 で固定したが，モデル

表 1　AUC-PR の手法ごとの比較（Kinship）.

	除いた正例の割合			
	0.1	0.3	0.5	0.7
RESCAL（式 (1)）	0.313	0.429	0.402	0.238
RESCAL-ALS ($\rho = 0.0$)	0.704	0.778	0.701	0.339
RESCAL-ALS ($\rho = 0.2$)	**0.741**	**0.790**	**0.754**	**0.692**
RESCAL-ALS ($\rho = 0.4$)	0.725	0.753	0.720	0.568

表 2　AUC-PR の手法ごとの比較（UMLS）.

	除いた正例の割合			
	0.1	0.3	0.5	0.7
RESCAL（式 (1)）	0.331	0.263	0.150	0.067
RESCAL-ALS ($\rho = 0.0$)	0.776	0.830	0.768	0.444
RESCAL-ALS ($\rho = 0.3$)	0.833	**0.917**	**0.878**	**0.745**
RESCAL-ALS ($\rho = 0.5$)	**0.897**	0.871	0.806	0.697

は，学習データ以外の一部 Fact を対象として Epoch ごとにランキングによる評価を行い，最も精度が高いものを使用した．実装は https://github.com/ibalazevic/TuckER に RESCAL を加えた[1]．実験結果を表 1, 2 に示す．どの設定でも提案手法が従来法より優れた分類精度を示した．

4.2　学習範囲による影響の分析

表 1, 2 では従来法と比較して PU なしの ALS ($\rho = 0.0$) も高い分類精度を示している．全 Triple を学習する ALS に対し，従来法は Fact と負例の一部のみ学習している．この違いが分類に与える影響を調査するため，従来法が学習において負に分類する範囲を Negative Area，学習を行わない範囲を Unlearned Area として，それぞれの分類精度を AUC-PR で評価した（図 4）[2]．二つの領域における正例と負例の比率を合わせるため，Unlearned Area で負例をランダムにサンプリングしテストした．従来法では，限られた範囲の学習によって Triple 全体における Fact の補完が可能であるとされてきたが，Unlearned Area ではほとんど分類ができていない．ALS での分類精度はこれと比較して高い．これは ALS が Unlearned Area の学習を行っているので直感的だが，それでも値が高くないのは，これらの Triple は近傍に正例を持たないため学習が困難だからだと考えられる．また Negative Area でも ALS の方が分類精度が高く，Unlearned Area の学習が特徴表現を通じ Fact の補完に重要な情報を与えていると考えられる．

1)　TuckER モデルでも同様の実験を行ったが，分類精度は RESCAL と同等であった．
2)　表 1 の除いた正例の割合 0.3 で表示したモデルを使用．

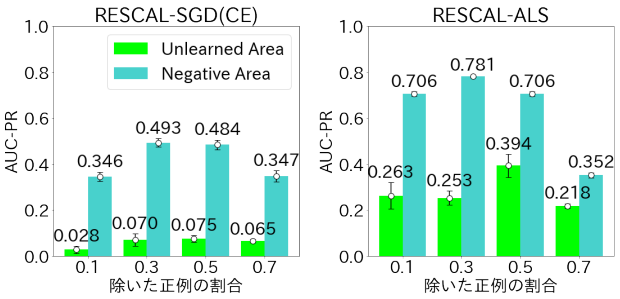

図 4　目的関数式 (1) が負例として学習する Triple（Negative Area ）と学習をしない Triple（Unlearned Area）を分け，AUC-PR で評価．左は (1) に基づく従来法，RESCAL-ALS（$\rho = 0.0$）による学習．

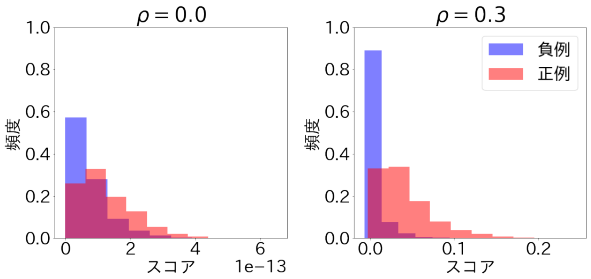

図 5　Fact を 0.8 除いた Kinship における正例負例のスコアの分布．PU なし $\rho = 0.0$（左）と PU あり $\rho = 0.3$（右）．

4.3　PU 学習が与える影響の分析

PU 学習なしの ALS に対して PU ありの ALS が高い分類精度を示しており，特に正例を除いた割合が高い時その差が大きい．PU の影響を調査するため，正例を 0.8 の割合除いて，学習に与えてない正例と負例のスコアの分布を $\rho = 0.0$ と $\rho = 0.3$ で比較した（図 5）[3]．$\rho = 0.0$ では Fact 以外を全て負に分類するためスコアが非常に低い値に密集しており正例と負例のスコアに差を持たせることができておらず分類精度が低い（AUC-PR=0.075）．一方 $\rho = 0.3$ ではスコアの密集が防がれ正例と負例のスコアに差が生まれており，分類精度が向上している（AUC-PR=0.378）．

4.4　実験まとめ

従来法は学習範囲が局所的という問題があり，分類精度が非常に低いのに対し，ALS はこの問題を回避している．また PU 学習を用いることで ALS による埋め込みの分類精度が向上した．一方 PU 学習の理論に適合しない結果もあり，より知識グラフに適した PU 定式化は今後の課題である．

3)　引き続き次元数 100，正則化項 $\lambda = 0.0001$ を使用．

参考文献

[1] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, p. 809–816, Madison, WI, USA, 2011. Omnipress.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 26, pp. 2787–2795. Curran Associates, Inc., 2013.

[3] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[4] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 213–220, New York, 2008. Association for Computing Machinery.

[5] Qinghua Guo Wenkai Li and Charles Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 49, No. 2, pp. 717–725, 2011.

[6] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 1196–1204. Curran Associates, Inc., 2013.

[7] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit S Dhillon. Pu learning for matrix completion. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 2445–2453. JMLR.org, 2015.

[8] Yanjie Wang, Rainer Gemulla, and Hui Li. On multi-relational link prediction with bilinear models. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 4227–4234. AAAI Press, 2018.

[9] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[10] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing yago: Scalable machine learning for linked data. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, p. 271–280, New York, NY, USA, 2012. Association for Computing Machinery.

[11] Théo Trouillon, Christopher R. Dance, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. *CoRR*, Vol. abs/1702.06879, , 2017.