

知識ベース埋め込みのためのペアワイズ積 L1 正則化

真鍋 陽俊^{*a} 林 克彦^b 新保 仁^a
^a 奈良先端科学技術大学院大学 情報科学研究科
^b 大阪大学 産業科学研究所
^{*}manabe.hitoshi.me@is.naist.jp

1 はじめに

YAGO や Freebase など大規模な知識ベースは質問応答や対話システムの構築に必要な不可欠である。こういった知識ベース中には、三つ組 (triplet) (s, r, o) の形式で表現された事実 (fact) が格納されており、この三つ組は二つのエンティティ (entity) s, o (それぞれ subject, object とよばれる) の間に関係 (relation) r が成り立つことを表している。既存の知識ベースは非常に大規模ではあるが、欠損している事実も数多く存在すると報告されている [3]。知識ベース補完 (knowledge base completion) は、こういった欠けている事実を自動的に発見することを目的としたタスクである。近年は、このタスクに対して、ベクトル埋め込みに基づく手法が盛んに提案されている。これらの手法では、各エンティティと関係をベクトル空間上に埋め込み、与えられた三つ組 (s, r, o) が事実としての妥当かどうかの度合いを (実数値として) 評価するスコアリング関数 (scoring function) $\phi(s, r, o)$ を、埋め込み空間上のベクトル演算を用いて定義する。

一般に subject と object が交換不可能な関係 (例えば、親子関係 *parent_of*) が数多く存在することから、多くの埋め込みベースの手法は、エンティティに対して対称性を満たさないスコアリング関数 ϕ (すなわち、あるエンティティ対 e_1, e_2 と関係 r について $\phi(e_1, r, e_2) = \phi(e_2, r, e_1)$ が成り立たない) をも表現できるようなモデル化を行っている。

しかし、このような対称性の有無は、関係毎に異なる。例えば、*sibling_of* といった関係は対称関係である。つまり $(e_1, \textit{sibling_of}, e_2)$ が事実であれば $(e_2, \textit{sibling_of}, e_1)$ も事実であると考えられる。一方で *parent_of* のように $(e_1, \textit{parent_of}, e_2)$ が事実であれば $(e_2, \textit{parent_of}, e_1)$ は決して事実でないという関係 (非対称関係; asymmetric relation) も多数存在する。本稿ではこの観測に基づき、関係毎の対称度合いを自動的に調節するようなモデリング手法の提案を行う。より具体的には、複素埋め込み [4] をベースとし、各関係の対称度合いを調節するにあたり不必要なパラメータのスパース化を行う。この際、一般的な L1 正則化項を用いるのではなく、パラメータ間の依存性を考慮したペアワイズ積による L1 正則化項を提案する。

人工データによる実験から、実際に提案手法がより望ましく関係の対称度合いを捉えられていることを示し、また実データを用いた実験結果から提案手法が比較対象の手法よりも予測精度が高いことを示す。

2 背景

本稿では \mathbb{R} を実数の集合、 \mathbb{C} を複素数の集合、 i は虚数単位とし、 $[\mathbf{v}]_j$ はベクトル \mathbf{v} の j 番目の要素、 $[\mathbf{M}]_{jk}$ は行列 \mathbf{M} の (j, k) 要素を表す。ある複素数、複素ベクトル、複素行列 \mathbf{Z} に対して $\bar{\mathbf{Z}}$ は共役複素数、 $\text{Re}(\mathbf{Z})$ と $\text{Im}(\mathbf{Z})$ それぞれ複素数の実部と虚部を表すものとする。

2.1 複素埋め込み

エンティティの集合を \mathcal{E} 、関係の集合を \mathcal{R} とする。知識ベース補完タスクでは未知の三つ組 (s, r, o) が与えられた時にその三つ組が事実であるかの真偽値を予測する (ただし $s, o \in \mathcal{E}, r \in \mathcal{R}$)。この予測器を構築するための典型的な手法として、スコアリング関数 $\phi(s, r, o; \Theta)$ を定義し、学習データから各エンティティと関係の潜在表現となるパラメータ集合 Θ を学習する埋め込みベースの手法が一般的である。

高性能な埋め込みベースの手法の一つに、複素埋め込み (complex embedding; ComplEx) [4] がある。この手法では各エンティティと関係を複素ベクトルで表現し、次式によりスコアリングを行う。

$$\phi(s, r, o; \Theta) = \text{Re}(\mathbf{e}_s^T \text{diag}(\mathbf{w}_r) \bar{\mathbf{e}}_o) \quad (1)$$

ここで $\mathbf{e}_s, \mathbf{e}_o \in \mathbb{C}^d$ はそれぞれのエンティティ s, o に対応する複素ベクトルを表し、 $\mathbf{w}_r \in \mathbb{C}^d$ は関係 r に対応する複素ベクトルを表している。式 (1) において \mathbf{e}_o には複素共役が用いられている一方で \mathbf{e}_s では用いられていない。このモデリングによって、複素埋め込みは対称でないスコアリング関数の表現が可能になっている。

3 関係の対称度に応じたスコアリング関数の学習

3.1 複素埋め込みにおけるスコアリング関数の対称/歪対称性

知識ベースに含まれる大多数の関係は、対称もしくは非対称な関係である。例えば、WordNet のデータセットにある 18 種の関係は残らず対称 (4 種) もしくは非対称 (14 種) である。特に *born_in* のように subject と object になるエンティティが異なる「型」を持つ場合、その関係は非対称な性質を持つ。例えば (*Barack_Obama, born_in, Hawaii*) が事実なのであれば、(*Hawaii, born_in, Barack_Obama*) は事実ではない。

複素埋め込みによって定義されるスコアリング関数に着目すると、もしある関係 r の複素ベクトル \mathbf{w}_r が実数ならスコアリング関数は $\phi(s, r, o) = \phi(o, r, s)$ を満たし、任意の s, o の組み合わせに対して対称なスコアリングを実現することができる。一

方で複素ベクトル \mathbf{w}_r が純虚数となる場合、スコアリング関数は $\phi(s, r, o) = -\phi(o, r, s)$ を満たす (以下では、そのようなスコアリング関数を歪対称な関数と呼ぶ)。つまり、 \mathbf{w}_r における実部は関係 r に対するスコアリング関数の対称性に、虚部は歪対称性*1にそれぞれ対応している。しかし、従来の複素埋め込みでは、実部・虚部を表すパラメータの間に特別な制約は課されておらず、ゆえに関係毎の異なる対称度合いは考慮されていない。

3.2 スパース化による対称/歪対称性の学習

そこで本稿では L1 ノルムによるスパース化を用いて不必要なパラメータを削減する。しかし、通常の L1 ノルムの正則化では実部と虚部パラメータ間の独立性を暗に仮定しており、パラメータ間の依存性を考慮していない。例えば対称関係の場合、複素ベクトルは実部の大部分を非零値に保ちつつ虚部を 0 とすることが望ましい。そのためには実部・虚部パラメータ間の依存性を考慮した正則化が必要である。そこで提案手法では実部と虚部パラメータのペアワイズ積での L1 正則化を使い、より適切なスパース化を施す。複素埋め込みを以下の目的関数により学習を行う。

$$\min_{\Theta} \sum_{(s,r,o) \in \Omega} \log(1 + \exp(-y_{rso}\phi(s, r, o; \Theta))) + \lambda(\alpha R_1(\Theta) + (1 - \alpha)R_2(\Theta)) \quad (2)$$

Ω は学習データとなる三つ組の集合、 $y_{rso} \in \{-1, 1\}$ は三つ組 (s, r, o) の真偽値、 $\lambda \geq 0$ は正則化全体の強さ、 $\alpha \in [0, 1]$ は 2 つの正則化項 R_1 と R_2 の間のバランスを調整するハイパーパラメータである。2 つの正則化項は次のように定義される。

$$R_1(\Theta) = \sum_{r \in \mathcal{R}} \sum_{k=1}^d |\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)|, \quad (3)$$

$$R_2(\Theta) = \|\Theta\|_2^2. \quad (4)$$

従来の複素埋め込みでの目的関数とは異なり、我々の目的関数には正則化項 R_1 が含まれている。 R_1 は式 (3) からわかるように L1 ノルムに基づいている。しかしながら、このノルムは \mathbf{w}_r の実部と虚部の積からなるベクトルに対して計算されており、単純にパラメータベクトル (この場合は \mathbf{w}_r) の L1 ノルムを用いる通常の L1 正則化とは異なる。各項 $|\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)|$ は、 $\operatorname{Re}([\mathbf{w}_r]_k)$ と $\operatorname{Im}([\mathbf{w}_r]_k)$ のいずれか一方が 0 であれば最小値 0 を取るため、これらのうち一方のみを 0 とすることを促進する。 \mathbf{w}_r の実部がスコアリング関数の対称性を司り、虚部が歪対称性を司ることを考慮すると、 R_1 が我々の意図した、各関係の対称度合いの自動調節に適したモデルを実現していることがわかる。

*1 ただし「歪対称なスコアリング関数」は、一般の「非対称関係」にとっての理想的なスコアリング関数とはならないことに注意されたい。非対称関係においては (e_1, r, e_2) が成り立たないからといって、 (e_2, r, e_1) が成り立つとは限らないが、歪対称なスコアリング関数はこれを暗に示している。

3.3 MAP 推定との関連

MAP 推定では事後確率を最大化するようにモデルパラメータ $\hat{\Theta}$ を点推定する手法である。

$$\hat{\Theta} = \operatorname{argmax}_{\Theta} \log p(\Theta | \mathcal{D}) = \operatorname{argmax}_{\Theta} \log p(\mathcal{D} | \Theta) + \log p(\Theta), \quad (5)$$

\mathcal{D} は観測データを表す。上式の第一項は尤度関数であり、第二項はパラメータの事前分布を表している。

我々の目的関数である式 (2) は式 (5) の形式をした MAP 推定として見なすことができる。目的関数の第一項は尤度関数に対応しており、パラメータの事前分布を正則化項として考えることができる。

ここでは提案したペアワイズ積 L1 正則化がどのような事前分布を仮定しているか考える。議論の簡略化のため、L2 正則化項は無視し L1 正則化項のみに焦点を当てる。

以下では、 C, C', C'', \dots をある定数とする。我々の L1 正則化項は以下の事前分布を仮定していることと等価である。

$$p(\Theta) = \prod_{r \in \mathcal{R}} p(\mathbf{w}_r)$$

ここで

$$p(\mathbf{w}_r) = \prod_{k=1}^d C \exp\left(-\frac{|\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)|}{C'}\right).$$

つまり、 $\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)$ の事前分布が平均 0 のラプラス分布に従う。この等価性は次のように示すことができる。

$$\begin{aligned} \log p(\Theta) &= \log \prod_{r \in \mathcal{R}} p(\mathbf{w}_r) \\ &= \log \prod_{r \in \mathcal{R}} \prod_{k=1}^d C \exp\left(-\frac{|\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)|}{C'}\right) \\ &= -C'' \sum_{r \in \mathcal{R}} \sum_{k=1}^d |\operatorname{Re}([\mathbf{w}_r]_k) \cdot \operatorname{Im}([\mathbf{w}_r]_k)| + C'''. \end{aligned}$$

また通常の L1 正則化は以下の式で定義される。

$$R_{\text{std L1}}(\Theta) = \sum_{r \in \mathcal{R}} \sum_{k=1}^d (|\operatorname{Re}([\mathbf{w}_r]_k)| + |\operatorname{Im}([\mathbf{w}_r]_k)|) \quad (6)$$

MAP 推定として解釈すると、これは以下の事前分布を仮定していることと等価である。

$$\begin{aligned} p(\mathbf{w}_r) &= \prod_{k=1}^d p(\operatorname{Re}([\mathbf{w}_r]_k)) p(\operatorname{Im}([\mathbf{w}_r]_k)) \\ &= \prod_{k=1}^d C'''' \exp\left(-\frac{|\operatorname{Re}([\mathbf{w}_r]_k)|}{C''''}\right) \exp\left(-\frac{|\operatorname{Im}([\mathbf{w}_r]_k)|}{C''''}\right) \\ &= \prod_{k=1}^d C'''' \exp\left(-\frac{|\operatorname{Re}([\mathbf{w}_r]_k)| + |\operatorname{Im}([\mathbf{w}_r]_k)|}{C''''}\right) \quad (7) \end{aligned}$$

式 (7) の対数を取ると、式 (6) と対応することは明らかである。つまり、 $\operatorname{Re}([\mathbf{w}_r]_k)$ と $\operatorname{Im}([\mathbf{w}_r]_k)$ がそれぞれ平均 0 のラプラス分布に従い、同一次元に属していたとしても実部と虚部のバ

ラメータは独立であると仮定されていることがわかる。今回の対称/歪対称なスコアリング関数のように、もしパラメータ間で互いに依存性がある場合、この仮定は望ましくない。

3.4 Exclusive group lasso との関係

Exclusive group lasso (EGL) は各パラメータグループに対して $l_{1,2}$ ノルム制約を加えることでグループ内でパラメータ選択を行うスパース化手法のひとつである [2]。

式 (2) において $\alpha = 2/3$ の場合、正則化項は次の形となる。 $\alpha R_1(\mathbf{w}_r) + (1 - \alpha)R_2(\mathbf{w}_r) = (1/3) \sum_{k=1}^d (\text{Re}([\mathbf{w}_r]_k) + \text{Im}([\mathbf{w}_r]_k))^2$ 。この場合、右辺は $l_{1,2}$ ノルム制約を表しており、各次元毎に実部・虚部の値をグループ化した EGL の一種と見なすことができる。

3.5 学習アルゴリズム

上記の目的関数の最適化をするにあたり、確率的勾配降下法のような一般的なオンライン学習アルゴリズムではスパース解を得ることが難しい。そこで本稿では効率的にスパース解を導出する手法である Regularized Dual Averaging (RDA) [5] を元に最適化を行う。RDA では各時刻 t で損失関数に対するパラメータの劣勾配の平均値 $\bar{\mathbf{g}}^{(t)}$ を算出しながら最適化を行う手法である。今回は各時刻 t で算出する劣勾配は式 (2) の第 1 項と第 3 項に対して計算を行い、以下のようにパラメータ \mathbf{w}_r の更新を行う。

$$\begin{aligned} \text{Re}([\mathbf{w}_r]_k^{(t+1)}) &= \begin{cases} 0, & \text{if } |[\bar{\mathbf{g}}_r]_k^{(t)}| \leq \beta |\text{Im}([\mathbf{w}_r]_k^{(t)})|, \\ \gamma, & \text{otherwise,} \end{cases} \\ \text{Im}([\mathbf{w}_r]_k^{(t+1)}) &= \begin{cases} 0, & \text{if } |[\bar{\mathbf{g}}_r]_k^{(t)}| \leq \beta |\text{Re}([\mathbf{w}_r]_k^{(t)})|, \\ \gamma', & \text{otherwise,} \end{cases} \end{aligned}$$

$\beta = \lambda\alpha$ は定数であり、 $\mathbf{g}_r, \mathbf{g}'_r \in \mathbb{R}^d$ はそれぞれ関係 r における実部と虚部の劣勾配を表す。また、

$$\begin{aligned} \gamma &= -\eta t \left([\bar{\mathbf{g}}_r]_k^{(t)} - \beta |\text{Im}([\mathbf{w}_r]_k^{(t)})| \text{sign}([\bar{\mathbf{g}}_r]_k^{(t)}) \right), \\ \gamma' &= -\eta t \left([\bar{\mathbf{g}}'_r]_k^{(t)} - \beta |\text{Re}([\mathbf{w}_r]_k^{(t)})| \text{sign}([\bar{\mathbf{g}}'_r]_k^{(t)}) \right). \end{aligned}$$

であり、 η は学習率である。またスパース化を施さないエンティティの複素ベクトルについては通常のオンライン学習の枠組みで最適化を行う。

4 実験

4.1 人工データ

提案法が対称/歪対称なスコアリング関数を期待通りに学習できるか確認するために人工データによる実験を行った。人工データとして、50 種のエンティティと 3 種の関係からなる知識ベースを用意した。3 種の関係の内、一つは対称な関係 (sym)、二つ目は歪対称な関係 (anti)*2、最後はどちらでもない関係 (other) である。その知識ベースを学習/開発/テストデータに 8:1:1 の割合でランダムに分割し、提案法によるペアワイズ積の L1 正則化と一般的な L1 正則化の比較を行った。式 (2) における埋め込みの次元は $d = 50$ とし、 $\alpha = 1.0$, $\lambda = 0.05$, $\eta = 0.1$ とした。

*2 歪対称関係を、「 (e_1, r, e_2) が成り立つ $\Leftrightarrow (e_2, r, e_1)$ が成り立たない」をみたす関係と定義する。

表 1: 人工データにおける分類精度 (%)

Models	sym	anti	other	all
ComplEx w/ std L1	89.8	92.1	65.7	81.5
ComplEx w/ mul L1	93.5	94.4	65.3	83.3

表 2: FB15K と WN18 のデータセットの統計量

Dataset	E	R	#train	#valid	#test
FB15k	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000

図 1 に実際に学習された各関係のベクトルを示す。これらからわかるように我々の提案法では期待したスパース化が行われている。すなわち、対称関係では虚部、歪対称関係では実部の値がスパース化されている。

表 1 にテストデータの三つ組に対する分類精度を示す。これは与えられた三つ組の真偽値を予測するタスクであり、出力スコアに対して sign 関数を適用することでシステムの真偽値予測を決定する。提案法 ('ComplEx w/ mul L1') が一般的な L1 正則化手法 ('ComplEx w/ std L1') より対称、歪対称関係において高精度な予測を実現できている事がわかる。

4.2 実データ: WN18, FB15K

実データでの実験として、WordNet (WN18) と Freebase (FB15K) を用いて提案法と他の手法と比較を行う。各データセットの統計量を表 2 に示す。

評価は先行研究にならい、各テストサンプル (s, r, o) に対して、object o (または subject s) を全てのエンティティ e で置き換えて、 (s, r, e) (または (e, r, o)) のスコアを計算する。そして、そのスコアに基づいて降順にソートし、平均逆順位 (MRR: mean reciprocal rank) とトップ N に正解が含まれる割合 ("Hits @ N ") を求めて、"raw" または "filtered" の設定で評価する。"filtered" では訓練、開発、テストデータ上の $e \neq o$ (または $e \neq s$) となる (s, r, e) (または (e, r, o)) を全て削除して評価し、"raw" では削除を行わず評価する。

モデル選択は開発データに対して filtered MRR の値が最大となるように行う。また、ハイパーパラメータは次の範囲でグリッドサーチにより決定する。 $\lambda \in \{0.01, 0.001, 0.0001, 0\}$, $\alpha \in \{0, 0.3, 0.5, 0.7, 1.0\}$, $\eta \in \{0.1, 0.05\}$ 。学習中の学習率 η は Adagrad [1] により調整を行った。最大で 500 epoch の学習を行い、ベクトルの次元数は 200 とした。また、負例サンプリング数は学習データに含まれる三つ組 1 個につき FB15K では 10、WN18 では 5 とした。

表 3 に結果を示す。提案手法と一般的な L1 正則化を加えた手法、従来の複素埋め込みを含めたいくつかのベースライン手法で比較を行った。比較手法において、DistMult 以外は非対称なスコアリング関数に基づく手法となっている。

我々の提案手法は、ほとんどの評価指標において、他の手法を上回るか、もしくは最高精度を出した手法と同等の性能を示した。また特に、FB15K のデータセットにおいて顕著な改善が見られた。

データセットによる改善度の違いは、各データセット中に

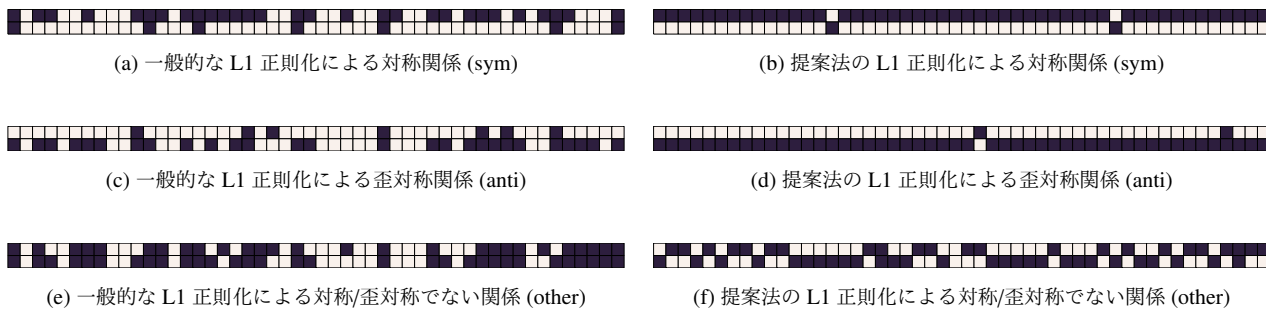


図 1: 人工データで学習された各関係の複素ベクトルの可視化。各列が複素ベクトルのそれぞれの要素を表しており、上部と下部はそれぞれ実部と虚部に対応している。黒色のセルは非ゼロの値を表す。

表 3: WN18 と FB15k での実験結果: (Filtered / raw) MRR と filtered Hits@{1, 3, 10} (%). *は [4] から引用した結果である。

Models	WN18					FB15k				
	MRR		Hits@			MRR		Hits@		
	Filter	Raw	1	3	10	Filter	Raw	1	3	10
TransE*	45.4	33.5	8.9	82.3	93.4	38.0	22.1	23.1	47.2	64.1
DistMult*	82.2	53.2	72.8	91.4	93.6	65.4	24.2	54.6	73.3	82.4
HolE*	93.8	61.6	93.0	94.5	94.9	52.4	23.2	40.2	61.3	73.9
ComplEx*	94.1	58.7	93.6	94.5	94.7	69.2	24.2	59.9	75.9	84.0
ComplEx ($\alpha = 0$)	94.3	58.2	94.0	94.6	94.8	69.5	24.2	59.8	76.9	85.0
ComplEx w/ std L1	94.3	57.9	94.0	94.5	94.8	71.1	25.5	61.8	78.3	85.6
ComplEx w/ mul L1	94.3	58.5	94.0	94.6	94.9	73.3	25.8	64.3	80.3	86.8

表 4: WN18 で学習データを半分にした場合の結果

Models	MRR		Hits@		
	Filter	Raw	1	3	10
ComplEx ($\alpha = 0$)	48.3	32.9	47.4	48.7	49.8
ComplEx w/ std L1	48.2	33.4	47.2	48.7	50.2
ComplEx w/ mul L1	49.0	34.6	47.7	49.7	51.2

含まれる出現頻度の少ない関係の割合によると推測される。FB15Kに含まれる 1,345 個の関係のほとんどは、学習データ数が十数個と非常に少ない。一方で WN18 に含まれる 18 関係のほとんどは十分な学習データが存在する構成となっている。スパース化は特に学習データの数が少ない場合に有効な手法として知られており、これが上述の性能改善の違いにつながったと考えられる。

この仮説を確認するために、提案法の効果が小さかった WN18 の学習データを半分のサイズに減らし、同様の設定で追加実験を行った。その結果を表 4 に示す。期待したように本来の WN18 より性能差は大きくなり、どの評価指標においても提案手法が最高精度となっている。

5 まとめ

本稿では複素埋め込みを基礎とし、各関係に応じて自動的にスコアリング関数の対称性を調整する正則化項を提案した。実験結果から、提案手法が従来の複素埋め込みの性能を上回ることを示した。

今後の研究の方向性としては、我々の手法により適した負例サンプリング法の提案が考えられる。なお本稿で用いたソース

コードは公開予定である。^{*3}

参考文献

- [1] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011.
- [2] Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via $\ell_{1,2}$ -norm. In *Advances in Neural Information Processing Systems 27*, pages 1655–1663, 2014.
- [3] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.
- [4] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning (ICML '16)*, pages 2071–2080, 2016.
- [5] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 22 (NIPS '09)*, 2009.

^{*3} <https://github.com/mana-ysh/symmetry-learning-kgc>