

能動学習による 多関係データセットの構築

梶野 洸¹, 岸本 章宏², Adi Botea²,
Elizabeth Daly², Spyros Kotoulas²

1. 東京大学

2. IBM Research - Ireland

手作業での多関係データセット構築支援という新たな枠組みを提案

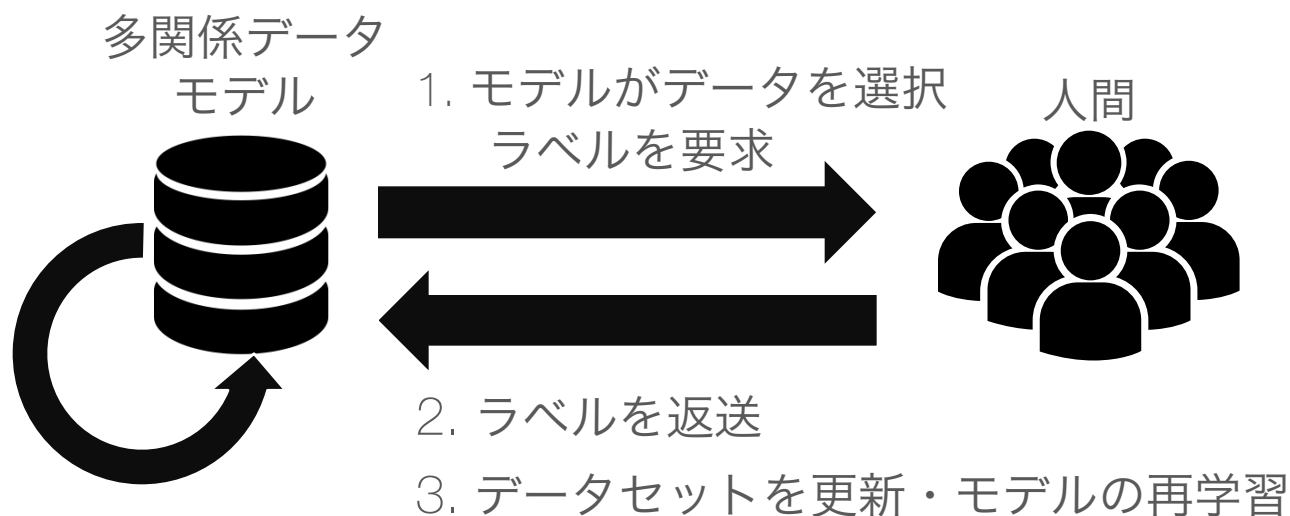
■ 研究対象: 手作業での多関係データセット構築

テキストからの自動抽出が困難なデータが存在

Q: 効率的に構築するにはどうしたらよいか?

■ 提案: 能動学習 + 多関係データ学習

能動学習により人間へのクエリ数を削減



■ 目次

□ 問題設定:

- 多関係データとその応用
- 2つの定式化:
 - データセット構築問題
 - 予測モデル構築問題

□ 提案手法 (AMDC):

- 能動学習を用いた枠組み
- 多関係データ学習

□ 実験

■ 目次

□ 問題設定:

- 多関係データとその応用
- 2つの定式化:
 - データセット構築問題
 - 予測モデル構築問題

□ 提案手法 (AMDC):

- 能動学習を用いた枠組み
- 多関係データ学習

□ 実験

多関係データセットは、二値ラベル付き三つ組の集合

■ 多関係データ: 三つ組 + ラベル

□ 三つ組: $t = (i, j, k)$

- 主体: $i, j \in E$
- 関係: $k \in R$

□ ラベル:

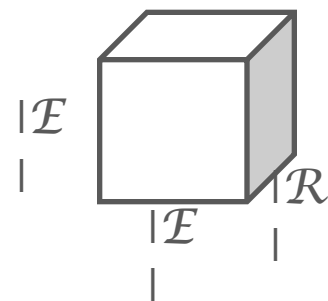
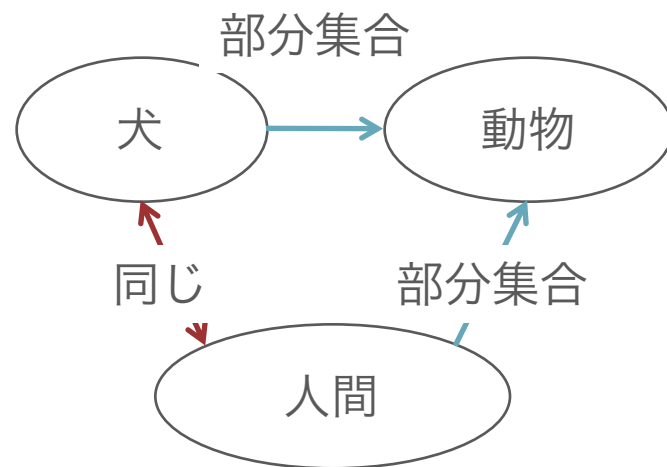
- t が**正例** $\Leftrightarrow$ 主体 i が主体 j に対して関係 k にある
- t が**負例** $\Leftrightarrow$ 主体 i が主体 j に対して関係 k にない

□ 多関係データセット: (Δ_p, Δ_n)

$$\Delta_p = \{t \in \Delta \mid t \text{ が正例}\}, \Delta_n = \{t \in \Delta \mid t \text{ が負例}\}$$

(仮定: $|\Delta_p| \ll |\Delta|$, ラベルなし三つ組も存在)

全三つ組の集合



人によるアノテーションが必要なデータが存在

■ 手作業によるデータセット構築の動機

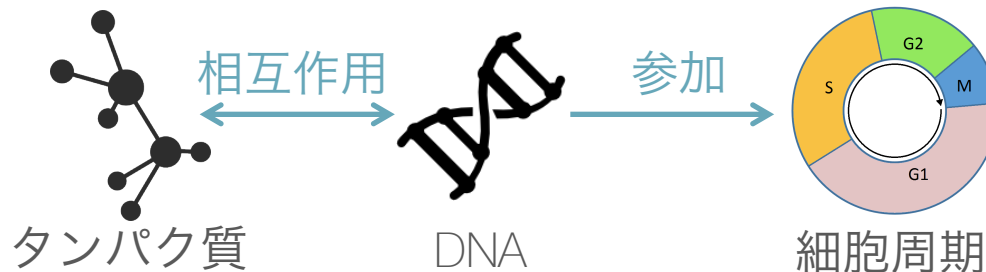
- 知識ベース: 人間の知識を多関係データで表現

データセット	正例の例
WordNet [Miller, 95]	(dog, canine, synset), (dog, poodle, hypernym)
ConceptNet [Liu+, 04]	(saxophone, jazz, UsedFor), (learn, knowledge, MotivatedByGoal)

ポイント: 常識的な知識は文書に明示的には現れない

→ 自動抽出が難しい

- 生物学のデータベース:



→ ラベル付けには実験が必要

データセットの用途に応じて2つの定式化をする

■ 2つの定式化

□ 入力:

- ・ 誤りなし
- ・ B 回クエリ可能

主体の集合 E , 関係の集合 R , オラクル $O: \Delta \rightarrow \{\text{正例}, \text{負例}\}$

□ 問題 1: データセット構築問題

目標: なるべく多くの正例を収集

□ 問題 2: 予測モデル構築問題

目標: 多関係データモデル $M: \Delta \rightarrow \mathbf{R}$ を学習 (出力: 正例らしさ)

※ モデルが欲しい場合、より直接的な定式化

■ 目次

□ 問題設定:

- 多関係データとその応用
- 2つの定式化:
 - データセット構築問題
 - 予測モデル構築問題

□ 提案手法 (AMDC):

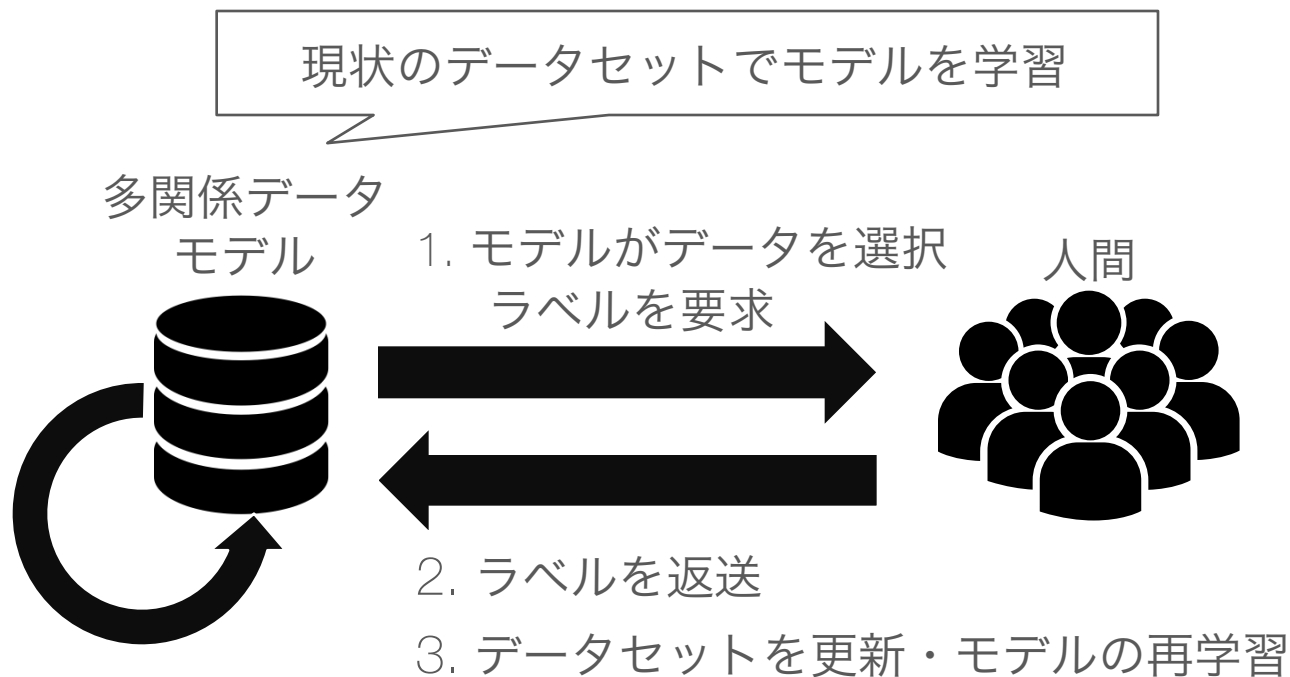
- 能動学習を用いた枠組み
- 多関係データ学習

□ 実験

AMDC はラベル取得と学習を B 回繰り返す

■ Active Multi-relational Data Construction (AMDC)

□ 概要:



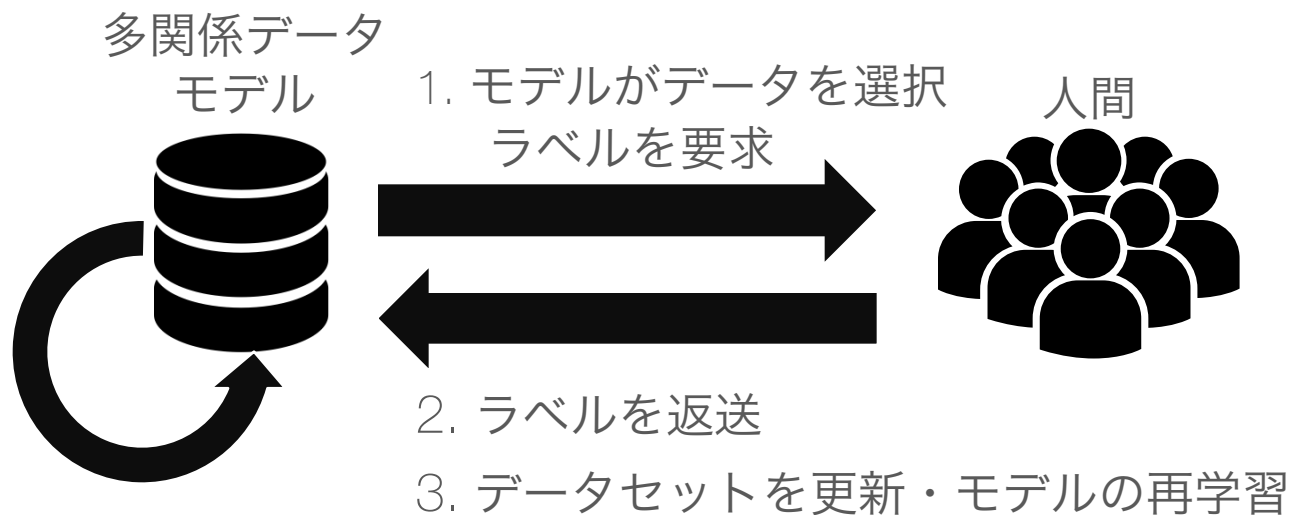
訓練データ (Δ_p, Δ_n)

AMDC はラベル取得と学習を B 回繰り返す

■ Active Multi-relational Data Construction (AMDC)

□ 概要:

AMDC は予測スコア s_t を計算可能
0より 大きい / 小さい $s_t \Leftrightarrow t$ を 正例 / 負例 と予測



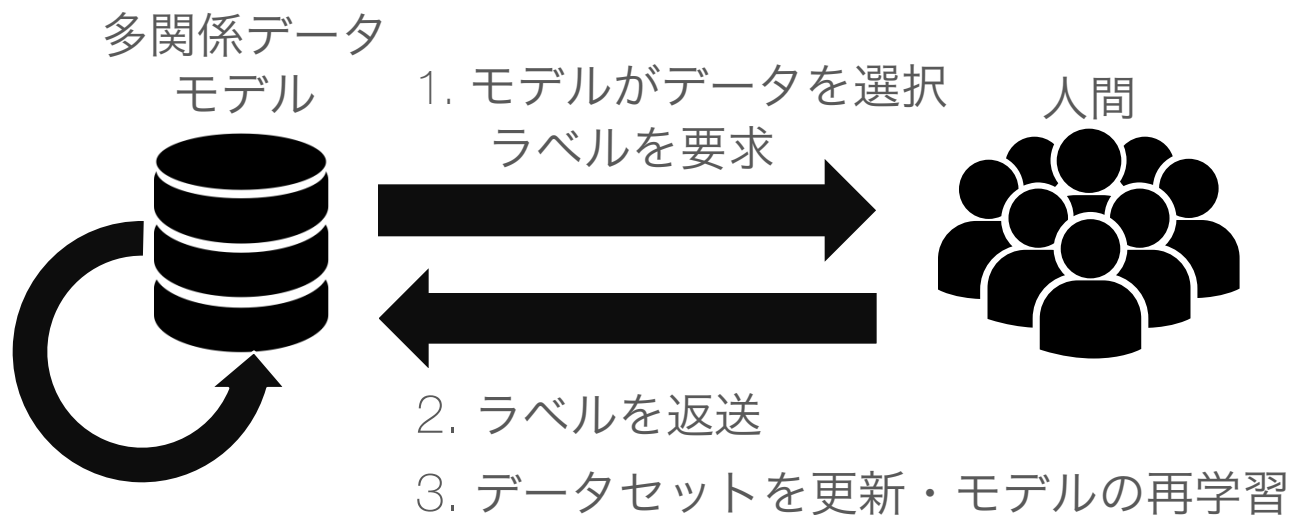
訓練データ (Δ_p, Δ_n)

AMDC はラベル取得と学習を B 回繰り返す

■ Active Multi-relational Data Construction (AMDC)

□ 概要:

予測スコア s_t を用いて クエリスコア q_t を計算
 q_t が小さい $\Leftrightarrow t$ のラベルが欲しい

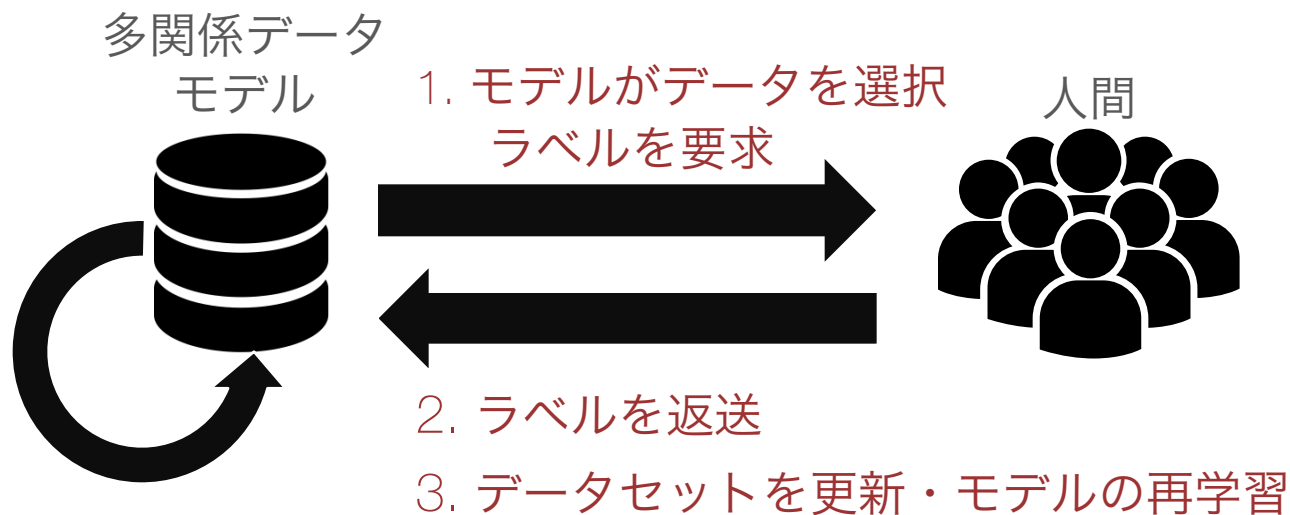


訓練データ (Δ_p, Δ_n)

AMDC はラベル取得と学習を B 回繰り返す

■ Active Multi-relational Data Construction (AMDC)

□ 概要:



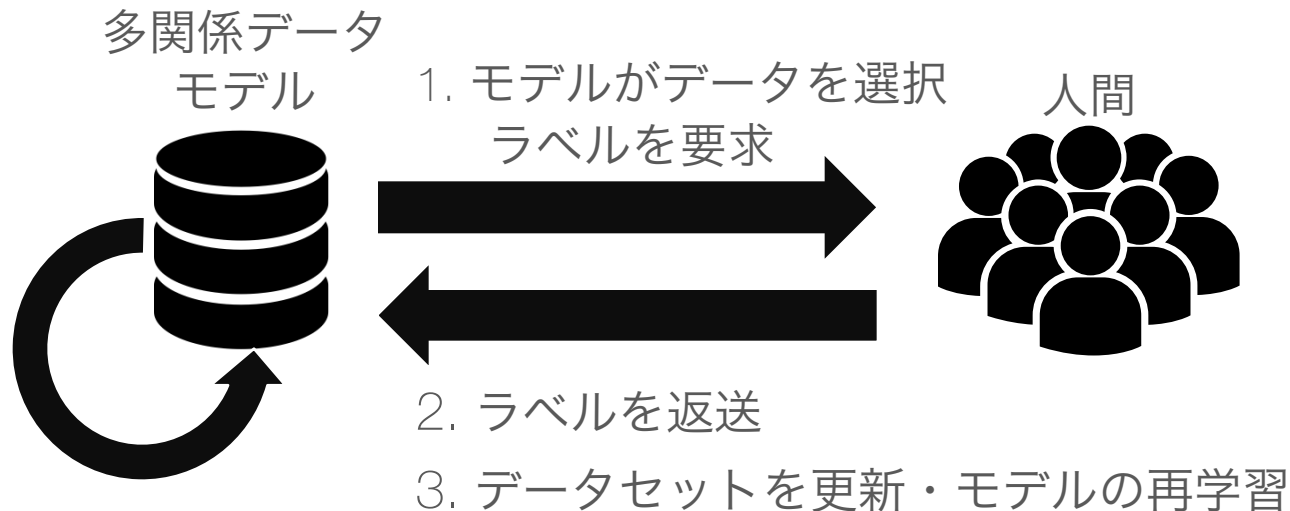
訓練データ (Δ_p, Δ_n)

AMDC を構成する2つのスコアを紹介する

■ Active Multi-relational Data Construction (AMDC)

□ 詳細:

- クエリスコア q_t
- 予測スコア s_t (多関係データモデル)

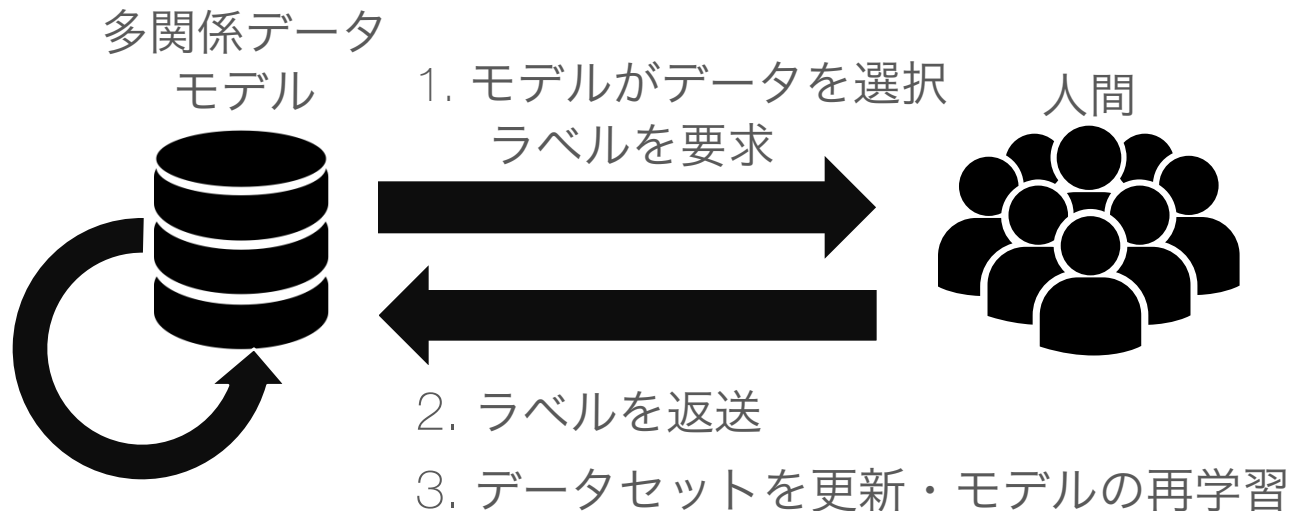


AMDC を構成する2つのスコアを紹介する

■ Active Multi-relational Data Construction (AMDC)

□ 詳細:

- クエリスコア q_t
- 予測スコア s_t (多関係データモデル)



2つの定式化それぞれに対しクエリスコアを設計

■ AMDC (1/2): クエリスコア q_t ($t \in \Delta$)

q_t が小さいほどデータセット構築に有用

入力: 予測スコア s_t

$s_t > 0$ (< 0) $\Leftrightarrow$ 正例 (負例) と予測

出力:

- 正例スコア (問題 1): $q_t := -s_t$

正例と予測された三つ組 (s_t が大) を選択

- 不確実スコア (問題 2): $q_t = |s_t|$

予測が不確実な三つ組 (境界上) を選択

※ クエリスコアを切り替えることで2つの問題設定に対応

2つの定式化それぞれに対しクエリスコアを設計

■ AMDC (1/2): クエリスコア q_t ($t \in \Delta$)

q_t が小さいほどデータセット構築に有用

入力: 予測スコア s_t

$s_t > 0$ (< 0) $\Leftrightarrow$ 正例 (負例) と予測

効果を実験で検証

出力:

- 正例スコア (問題 1): $q_t := -s_t$

正例と予測された三つ組 (s_t が大) を選択

- 不確実スコア (問題 2): $q_t = |s_t|$

予測が不確実な三つ組 (境界上) を選択

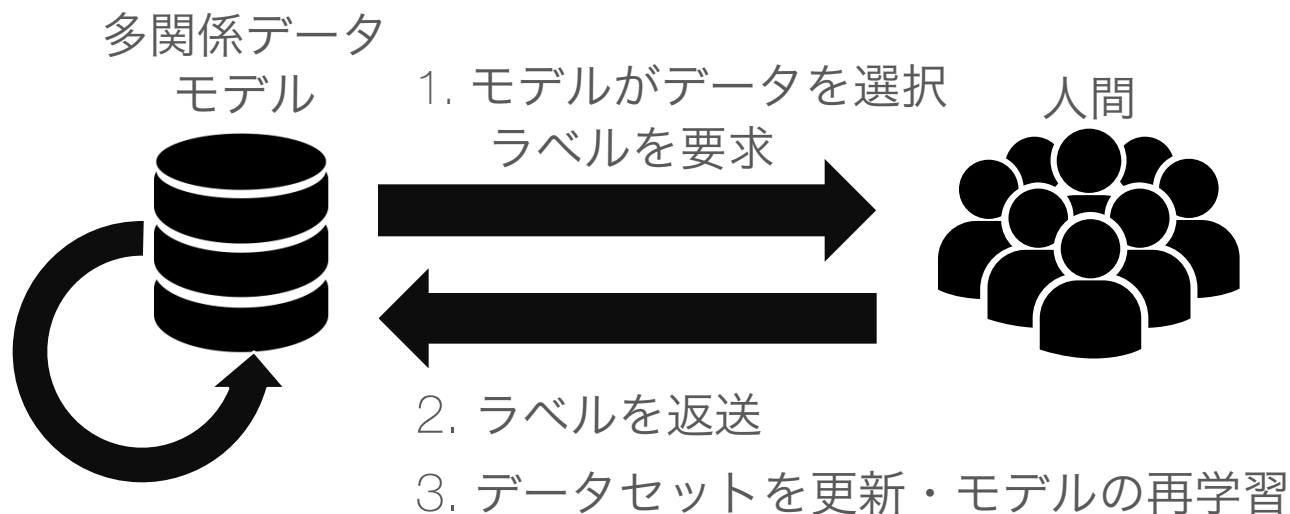
※ クエリスコアを切り替えることで2つの問題設定に対応

AMDC を構成する2つのスコアを紹介する

■ Active Multi-relational Data Construction (AMDC)

□ 詳細:

- クエリスコア q_t
- 予測スコア s_t (多関係データモデル)



過学習防止のために制約を導入

■ AMDC (2/2): 多関係データモデル

□ ベース: RESCAL [Nickel+, 11]

• モデル:

– $\mathbf{a}_i \in \mathbf{R}^D$: 主体 i の潜在表現 (ベクトル)

– $R_k \in \mathbf{R}^{D \times D}$: 関係 k の潜在表現 (行列 (対称とは限らない))

• 予測スコア: $s_t = \mathbf{a}_i^T R_k \mathbf{a}_j$

大きい / 小さい $s_t \Leftrightarrow t$ は 正例 / 負例

New

□ 新たな制約: $|\mathbf{a}_i| = 1, R_k = \text{回転行列}$

初期段階 (ラベル数が少ない) の学習を安定化

(→ 実験で検証)
18/29

ランキングベースの目的関数を利用

■ AMDC (2/2): 最適化問題

$$\min_{A,R} \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \sum_{\bar{t}_p \in \Delta \setminus \Delta_p} [\gamma - s_{t_p} + s_{\bar{t}_p}]_+ \quad \text{ } \boxed{s(\text{正例}) > s(\text{非正例})}$$

ランキングベースの目的関数を利用

■ AMDC (2/2): 最適化問題

$$\min_{A,R} \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \sum_{\bar{t}_p \in \Delta \setminus \Delta_p} [\gamma - s_{t_p} + s_{\bar{t}_p}]_+ \quad \text{ } s(\text{正例}) > s(\text{非正例})$$

毎反復で得られるラベルの
多くは負例

- ・ ラベル無しと負例を区別していない
- ・ 正例 / 負例 の境界が 0 となるとは限らない

ランキングベースの目的関数を利用

■ AMDC (2/2): 最適化問題

$$\min_{A,R} \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \sum_{\bar{t}_p \in \Delta \setminus \Delta_p} [\gamma - s_{t_p} + s_{\bar{t}_p}]_+ \quad \text{ } s(\text{正例}) > s(\text{非正例})$$
$$+ \frac{C_n}{|\Delta_n| |\Delta \setminus \Delta_n|} \sum_{t_n \in \Delta_n} \sum_{\bar{t}_n \in \Delta \setminus \Delta_n} [\gamma - s_{\bar{t}_n} + s_{t_n}]_+ \quad \text{ } s(\text{非負例}) > s(\text{負例})$$

・ ラベル無しと負例を区別していない

・ 正例 / 負例 の境界が 0 となるとは限らない

ランキングベースの目的関数を利用

■ AMDC (2/2): 最適化問題

$$\begin{aligned} \min_{A,R} & \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \sum_{\bar{t}_p \in \Delta \setminus \Delta_p} [\gamma - s_{t_p} + s_{\bar{t}_p}]_+ \quad \boxed{s(\text{正例}) > s(\text{非正例})} \\ & + \frac{C_n}{|\Delta_n| |\Delta \setminus \Delta_n|} \sum_{t_n \in \Delta_n} \sum_{\bar{t}_n \in \Delta \setminus \Delta_n} [\gamma - s_{\bar{t}_n} + s_{t_n}]_+ \quad \boxed{s(\text{非負例}) > s(\text{負例})} \\ & + C_e \left(\frac{1}{|\Delta_p|} \sum_{t_p \in \Delta_p} [\gamma' - s_{t_p}]_+ \quad \boxed{s(\text{正例}) > 0} + \frac{1}{|\Delta_n|} \sum_{t_n \in \Delta_n} [\gamma' + s_{t_n}]_+ \quad \boxed{s(\text{負例}) < 0} \right) \end{aligned}$$

- ・ ラベル無しと負例を区別していない

- ・ 正例 / 負例 の境界が 0 となるとは限らない

毎反復、複数のモデルをSGDで最適化し、最良のものを選択

■ AMDC (2/2): 最適化問題

$$\begin{aligned} \min_{A, R} & \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \sum_{\bar{t}_p \in \Delta \setminus \Delta_p} [\gamma - s_{t_p} + s_{\bar{t}_p}]_+ \quad \boxed{s(\text{正例}) > s(\text{非正例})} \\ & + \frac{C_n}{|\Delta_n| |\Delta \setminus \Delta_n|} \sum_{t_n \in \Delta_n} \sum_{\bar{t}_n \in \Delta \setminus \Delta_n} [\gamma - s_{\bar{t}_n} + s_{t_n}]_+ \quad \boxed{s(\text{非負例}) > s(\text{負例})} \\ & + C_e \left(\frac{1}{|\Delta_p|} \sum_{t_p \in \Delta_p} [\gamma' - s_{t_p}]_+ \quad \boxed{s(\text{正例}) > 0} + \frac{1}{|\Delta_n|} \sum_{t_n \in \Delta_n} [\gamma' + s_{t_n}]_+ \quad \boxed{s(\text{負例}) < 0} \right) \end{aligned}$$

□ アルゴリズム: SGD

□ パラメタ:

$$A = [\mathbf{a}_1 \ \dots \ \mathbf{a}_{|\mathcal{E}|}]^\top \in \mathbb{R}^{|\mathcal{E}| \times D}, R = [R_k]_{k \in \mathcal{R}} \in \mathbb{R}^{D \times D \times |\mathcal{R}|}$$

□ ハイパーパラメタ: $\gamma, \gamma', C_n, C_e, D$

毎反復、検証用データセットを用いて決定

毎反復、複数のモデルをSGDで最適化し、最良のものを選択

■ AMDC (2/2): 最適化問題

$$\min_{A,R} \frac{1}{|\Delta_p| |\Delta \setminus \Delta_p|} \sum_{t_p \in \Delta_p} \text{効果を実験で検証} + \boxed{s(\text{正例}) > s(\text{非正例})}$$

$$+ \frac{C_n}{|\Delta_n| |\Delta \setminus \Delta_n|} \sum_{t_n \in \Delta_n} \sum_{\bar{t}_n \in \Delta \setminus \Delta_n} [\gamma - s_{\bar{t}_n} + s_{t_n}]_+ \boxed{s(\text{非負例}) > s(\text{負例})}$$

$$+ C_e \left(\frac{1}{|\Delta_p|} \sum_{t_p \in \Delta_p} [\gamma' - s_{t_p}]_+ \underbrace{\boxed{s(\text{正例}) > 0}}_{\substack{\text{正例のスコアが} \\ \text{閾値} \gamma' \text{より大きい}}} + \frac{1}{|\Delta_n|} \sum_{t_n \in \Delta_n} [\gamma' + s_{t_n}]_+ \underbrace{\boxed{s(\text{負例}) < 0}}_{\substack{\text{負例のスコアが} \\ \text{閾値} \gamma' \text{より小さい}}} \right)$$

□ アルゴリズム: SGD

□ パラメタ:

$$A = [\mathbf{a}_1 \ \dots \ \mathbf{a}_{|\mathcal{E}|}]^\top \in \mathbb{R}^{|\mathcal{E}| \times D}, R = [R_k]_{k \in \mathcal{R}} \in \mathbb{R}^{D \times D \times |\mathcal{R}|}$$

□ ハイパーパラメタ: $\gamma, \gamma', C_n, C_e, D$

毎反復、検証用データセットを用いて決定

■ 目次

□ 問題設定:

- 多関係データとその応用
- 2つの定式化:
 - データセット構築問題
 - 予測モデル構築問題

□ 提案手法 (AMDC):

- 能動学習を用いた枠組み
- 多関係データ学習

□ 実験

3つの改良点の貢献を検証する

■ 実験

□ 目的: 3つの改良点の貢献を2つの問題設定で検証

- クエリスコア (vs. AMDC + random query)
- Neg AUC-loss (vs. AMDC - neg-AUC)
- RESCAL に加えた制約 (vs. AMDC - constraints)

□ データセット:

既に与えられているラベルを用いてオラクルをシミュレート

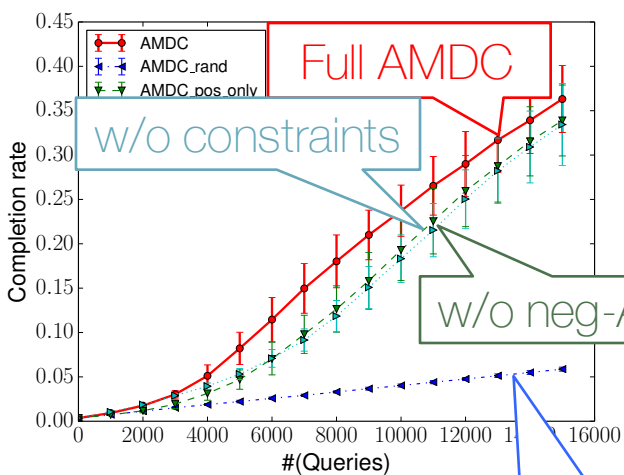
	#(主体)	#(関係)	#(正例)	#(負例)
Kinships [Denham, 73]	104	26	10,790	270,426
Nations [Rummel, 50-65]	125	57	2,565	8,626
UMLS [McCray,03]	135	49	6,752	886,273

AMDC はベースラインと比べ 2.4 – 19 倍の正例を収集

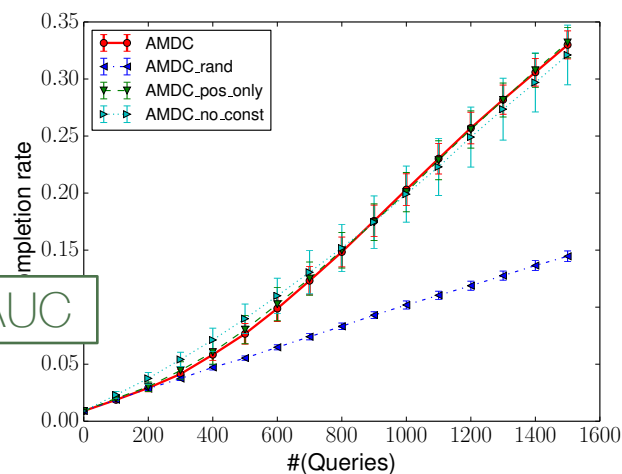
■ 実験 (1/2): データセット構築問題

評価指標: %(AMDC が収集した正例ラベル)

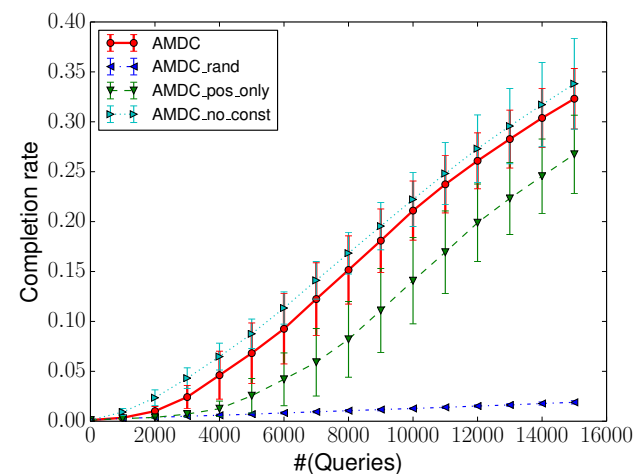
- クェリスコア: Random の 2.4 – 19 倍の正例を収集
- Neg AUC loss: 負例が多い場合に有効 (K, U)
- 制約: 性能が下がることもある



Kinships



Nations



UMLS

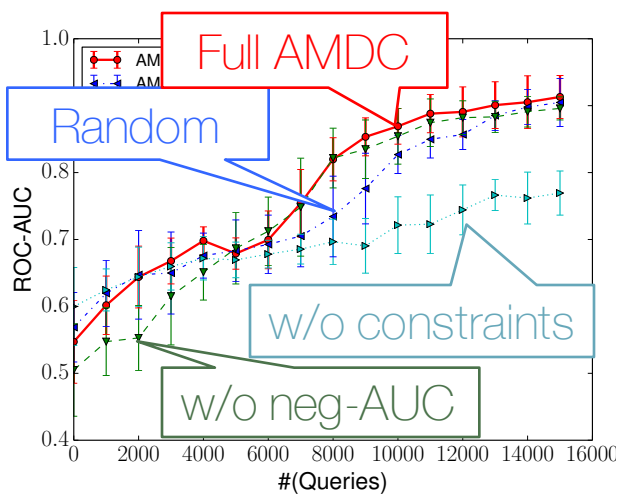
10 trials, $(Q, q) = (10^5, 10^3)$ ($(2 \times 10^3, 10^2)$ for Nations)

AMDC はほぼ常に最良の予測スコアを達成

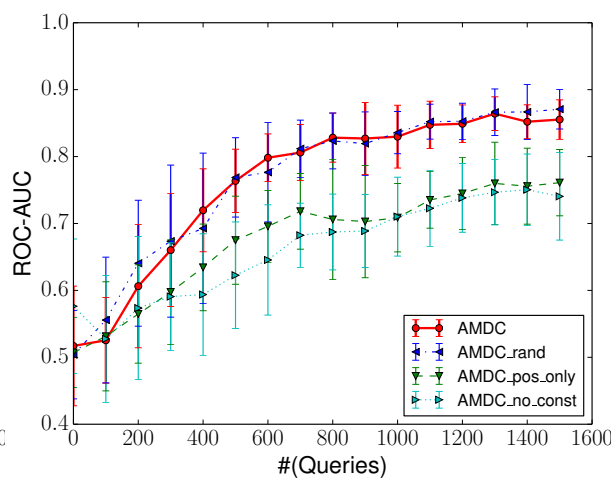
■ 実験 (2/2): 予測モデル構築問題

評価指標: ROC-AUC

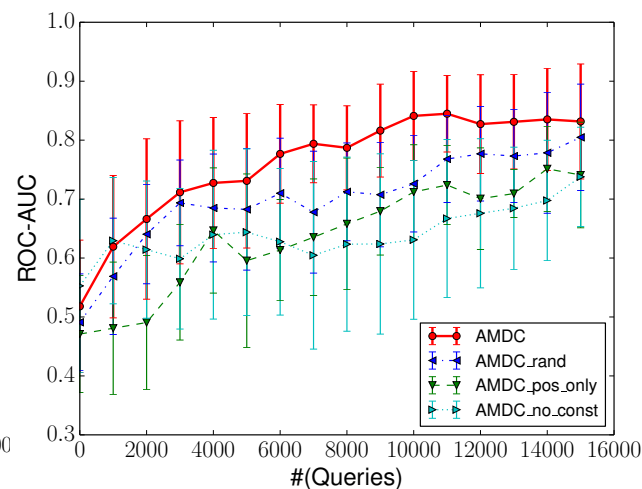
- クエリスコア: Random と比べ、有意に高いAUC (K, U)
- Neg AUC loss: どのデータでも有効
- 制約: 過学習を防ぎ、高い予測性能を達成



Kinships



Nations



UMLS

10 trials, $(Q, q) = (10^5, 10^3)$ ($(2 \times 10^3, 10^2)$ for Nations)

■ 結論

- 手作業での多関係データセット構築は未だに重要
 - ある種のデータは人によるアノテーションが必要
 - クラウドソーシングを通じて依頼可能
- AMDC = 能動学習 + 多関係データ学習

RESCAL を元に、データが少ない時の学習を安定化
- 3つの工夫により性能向上することを確認
 - 能動学習によりコストを大幅に削減
 - RESCALに加えた制約により過学習を防止
 - 負例が多い場合に neg AUC-loss が有効