[S3-P21]

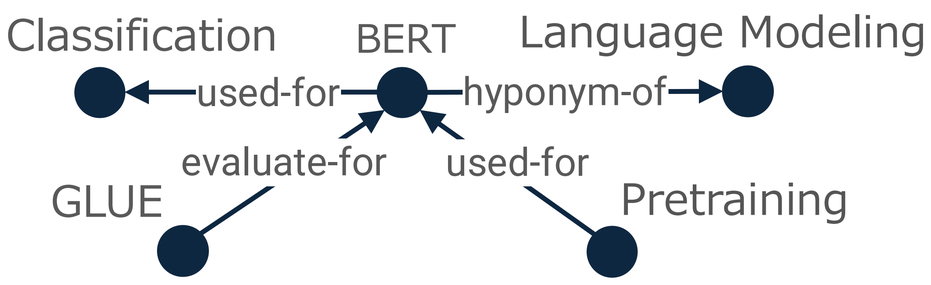
科学技術文献における知識グラフ補完を用いた効率的な知識グラフの作成

神野 倫行[1]　林 和樹[1]　坂井 優介[1]　上垣外 英剛[1]　渡辺 太郎[1]
(1. 奈良先端科学技術大学院大学/NAIST)

TL;DR: 知識グラフ補完による科学的発見の支援に向けて

- 背景：知識グラフの補完は、**新たな科学的知識の発見**や推薦に繋がることがある
- 課題：科学知識を含む訓練データの**手動作成**には、専門知識と**時間が必要**である
- 手法：**関係抽出モデル**を活用し訓練データの**自動構築**を試みる

背景: KG, KGCとは？



知識グラフ(Knowledge Graph)とは？
- 物同士の関係性を示したグラフ構造
- 関係は**トリプレット**で表現される
  - (head entity, relation, tail entity)
  - E.g. (BERT, used-for, classification)

知識グラフ補完(KG Completion)とは？
- KGから欠損している関係性を予測
  - 入力: (head entity, relation, ?)
  - 出力：tail entity【出力はランキング形式】

科学的知識を扱う理由
- 科学的知識の発見や推薦に繋がる可能性

課題: 訓練データの手動作成は高コスト

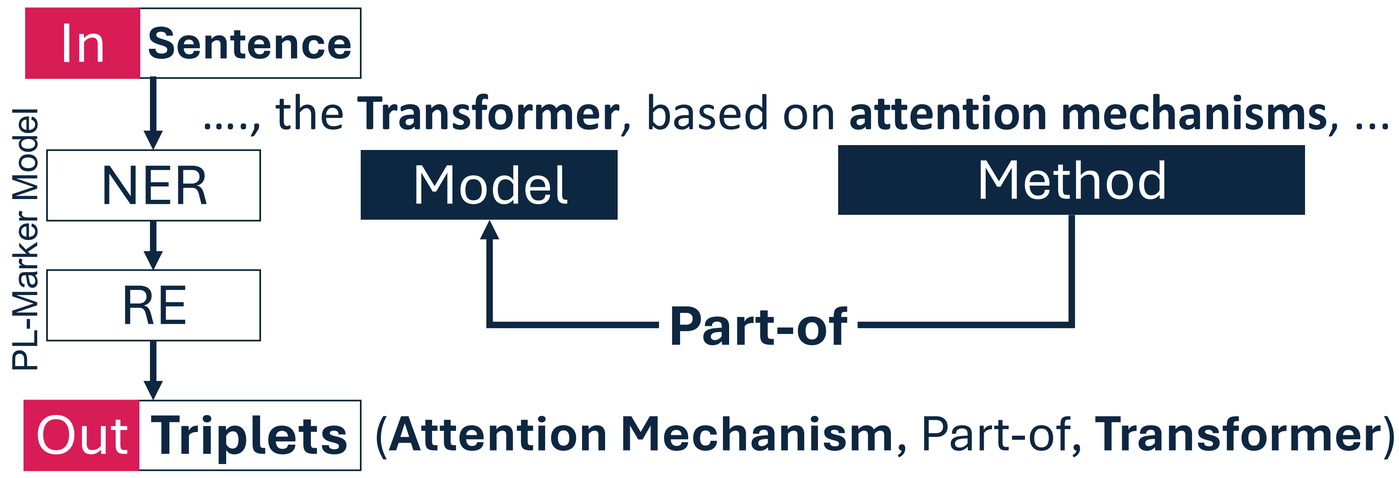
- **トリプレットの集合**がKGCモデルの**訓練データ**となる

訓練データサンプル

⋮
(knowledge graph completion, used-for, predicting missing link)
(bleu, evaluate-for, translation model)
⋮

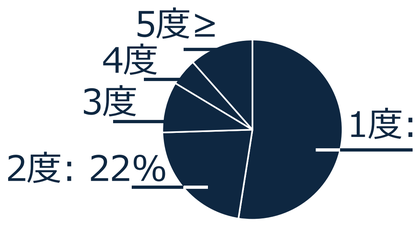
制作手法: 関係抽出モデルを使用

論文に関係抽出(Relation Extraction)を行い、訓練データ用のトリプレットを取得

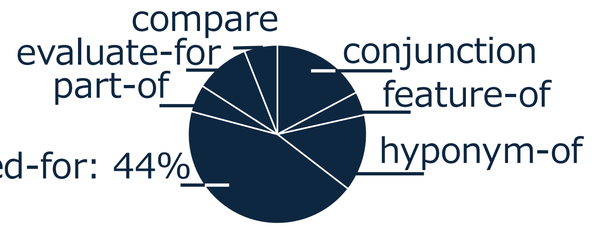


20万のエンティティと40万のトリプレットを取得

エンティティ頻度分布



関係区分分布



訓練データ 一部の放射状ツリー

予備実験: 当データセットで訓練されたTransEの性能
(出現頻度10以下をフィルタリングしたもの)

当データセット		
# of Training Triplets	# of Testing Triplets	Total # of Entities
264540	1000	36405

TransE			
Mean Rank	Hit@1	Hit@50	Hit@500
9180	0.011	0.366	0.581

今後の課題:

- 言語モデルを用いたKGCモデル用データセットの作成
- スパースなエンティティによる影響の調査
- より効果的なデータリーケージ予防手法の実装
- データセット内に含まれるノイズの軽減