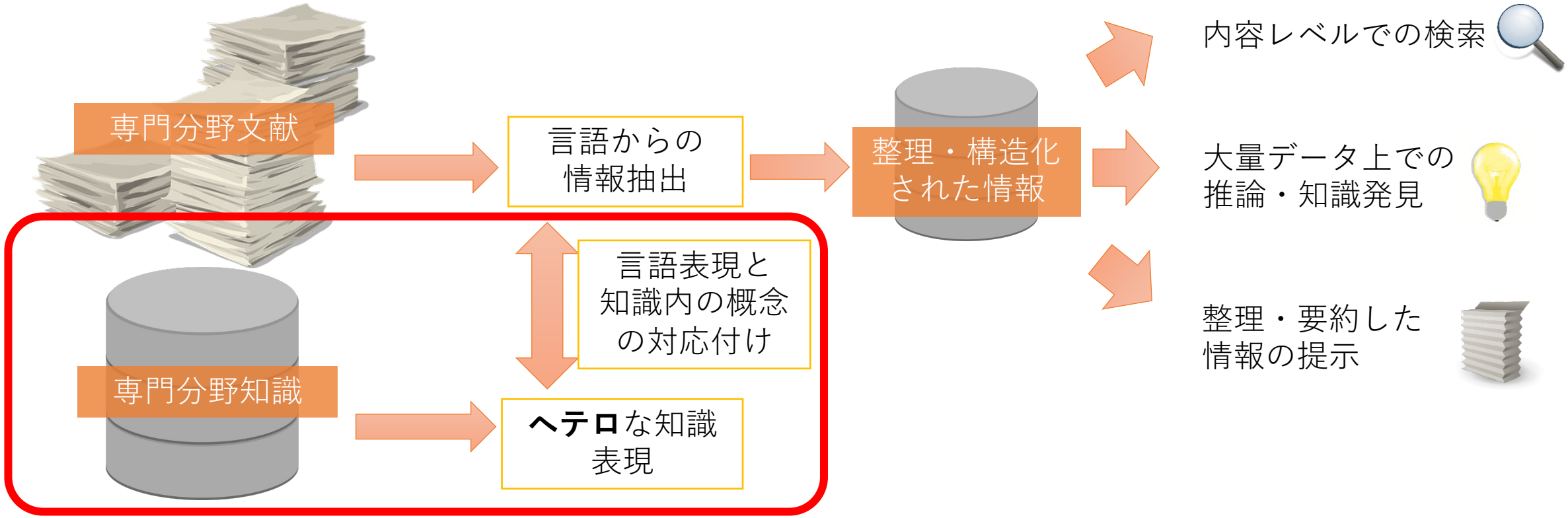
スマート情報研究センター
センター報告

ヘテロな専門分野知識に基づいた文献からの情報抽出

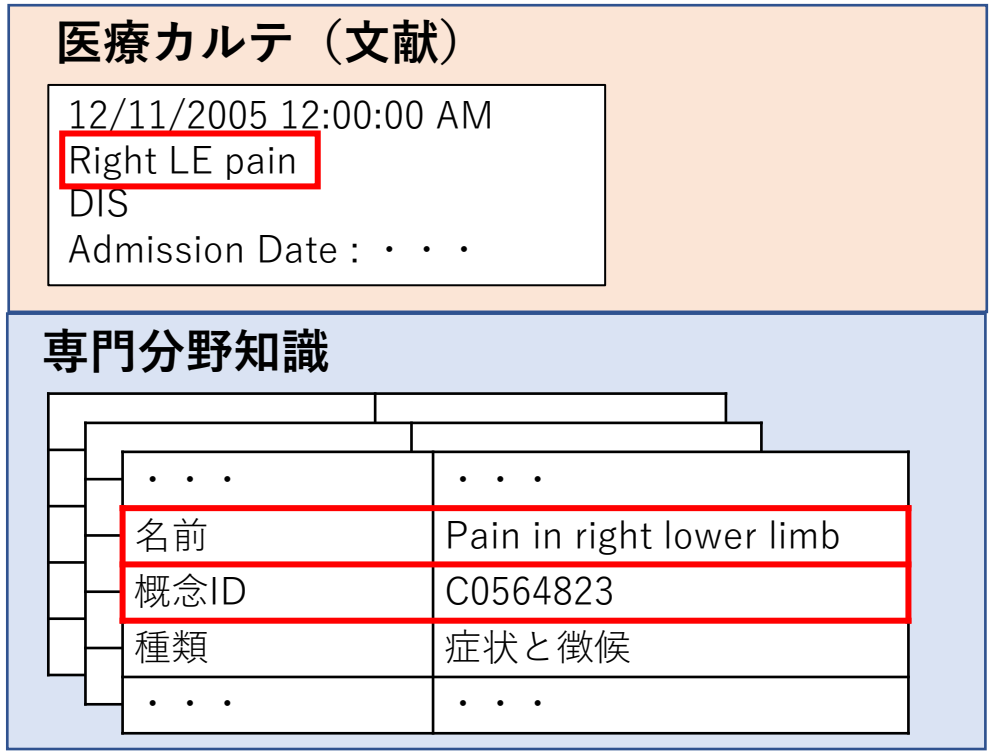
知能数理研究室 准教授 三輪誠

研究の概要

深層学習を基盤とした**ヘテロな専門分野知識に紐づいた**言語理解



テーマ 1: 医療分野における用語の言語表現と知識内の概念の対応づけ



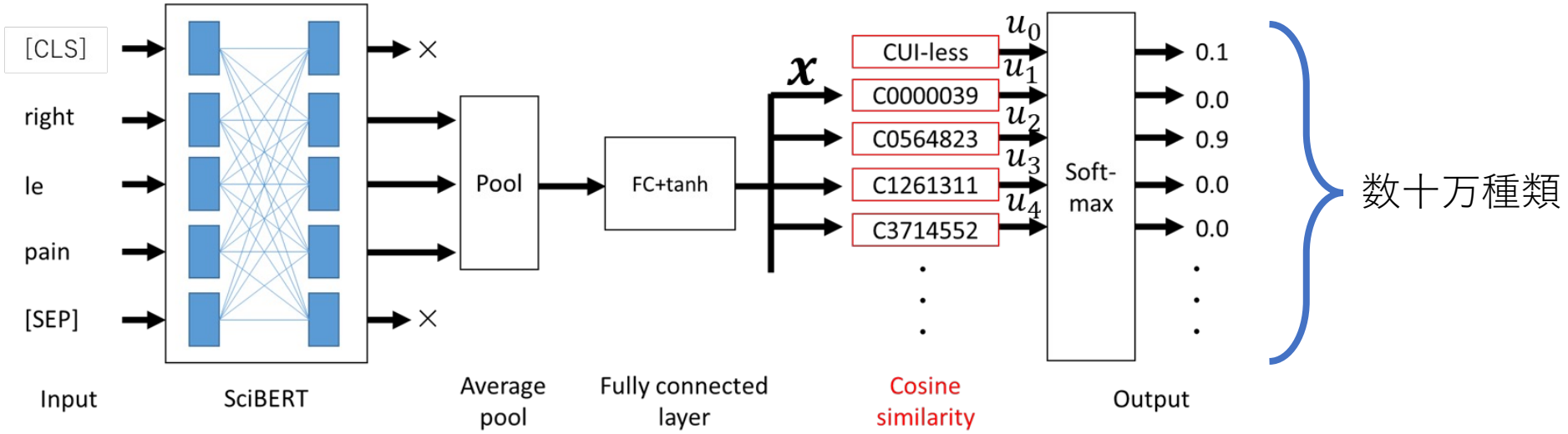
文献内の言語表現と専門分野知識内の概念は必ずしも簡単には対応が取れない

➔**言語表現**と**知識内の概念**（**数十万種類**）の対応づけ（**用語リンキング**）が必要

従来は辞書一致やルールによる言い換えにより対応

用語リンキングのための深層学習モデル

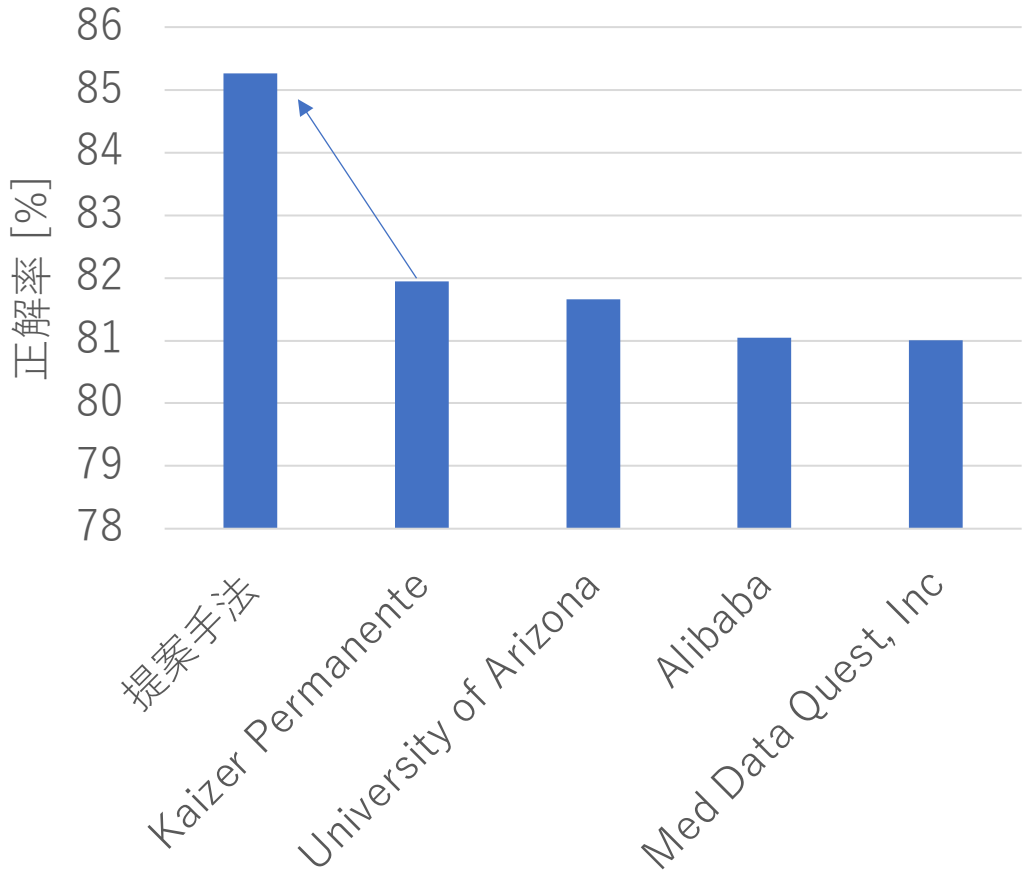
- 言語表現を数十万種類の知識内の概念のうちの一つに分類
- 言い換え辞書を追加の学習データとして利用してデータ増強
  - 言語表現と専門知識内の概念の対応づけ（数千）
  - **専門分野知識内の概念の言い換え（同義語）（数百万）**



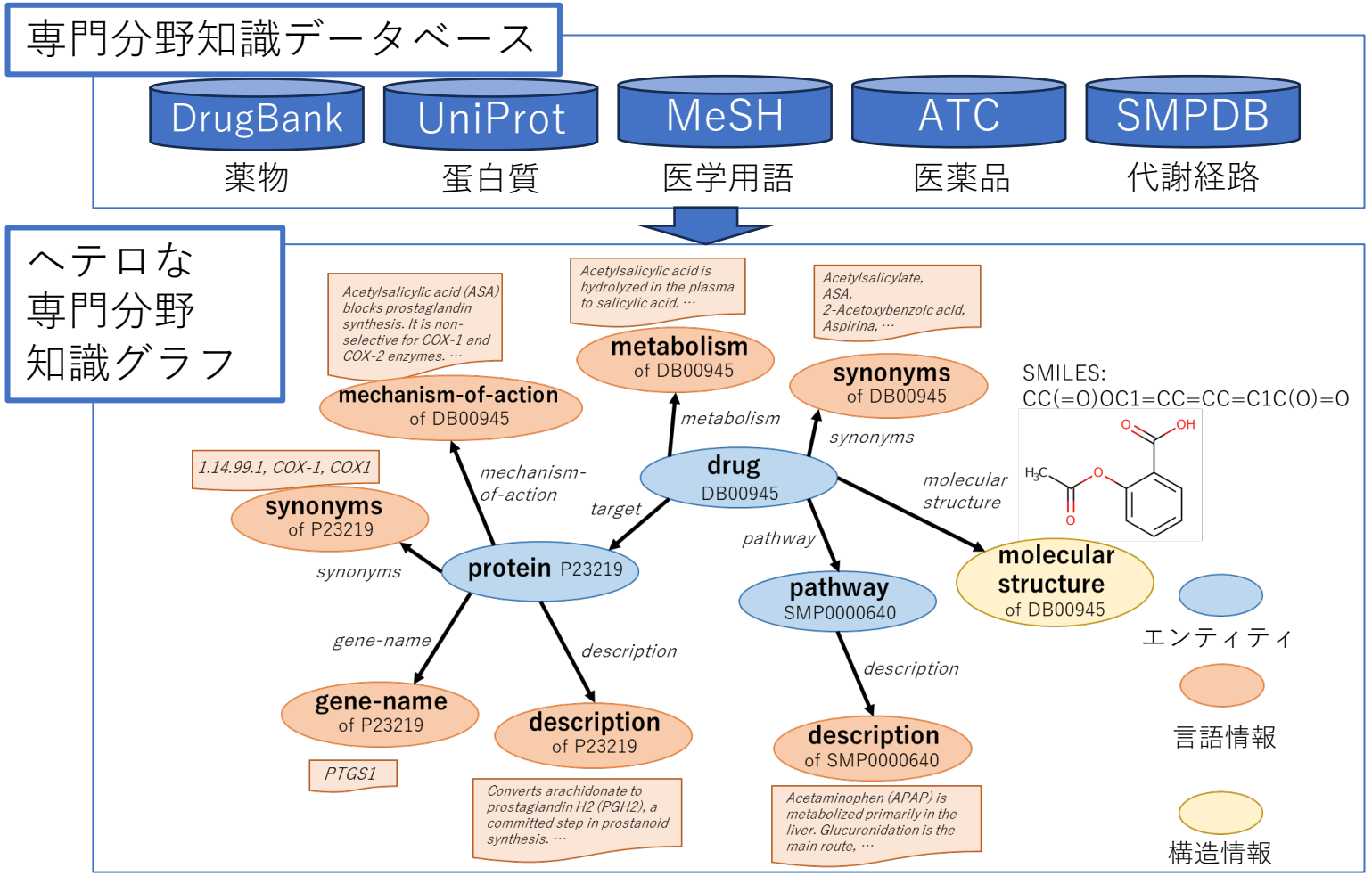
Tomoki Tsujimura, Makoto Miwa, and Yutaka Sasaki. Large-scale neural biomedical entity linking with layer overwriting. Journal of Biomedical Informatics, Vol. 143, 104433, 2023

用語リンキングモデルの性能

- 辞書マッチや人手の言い換えルールを利用する手法に比べ，**深層学習のみ**での用語リンキングによる性能向上
  - 正解率 81.94% ➔ 85.26%
- 深層学習により列挙や表現の難しい細かい言い換えのルールを大量のデータから学習



テーマ2: 薬学分野における専門分野知識の情報抽出への利用



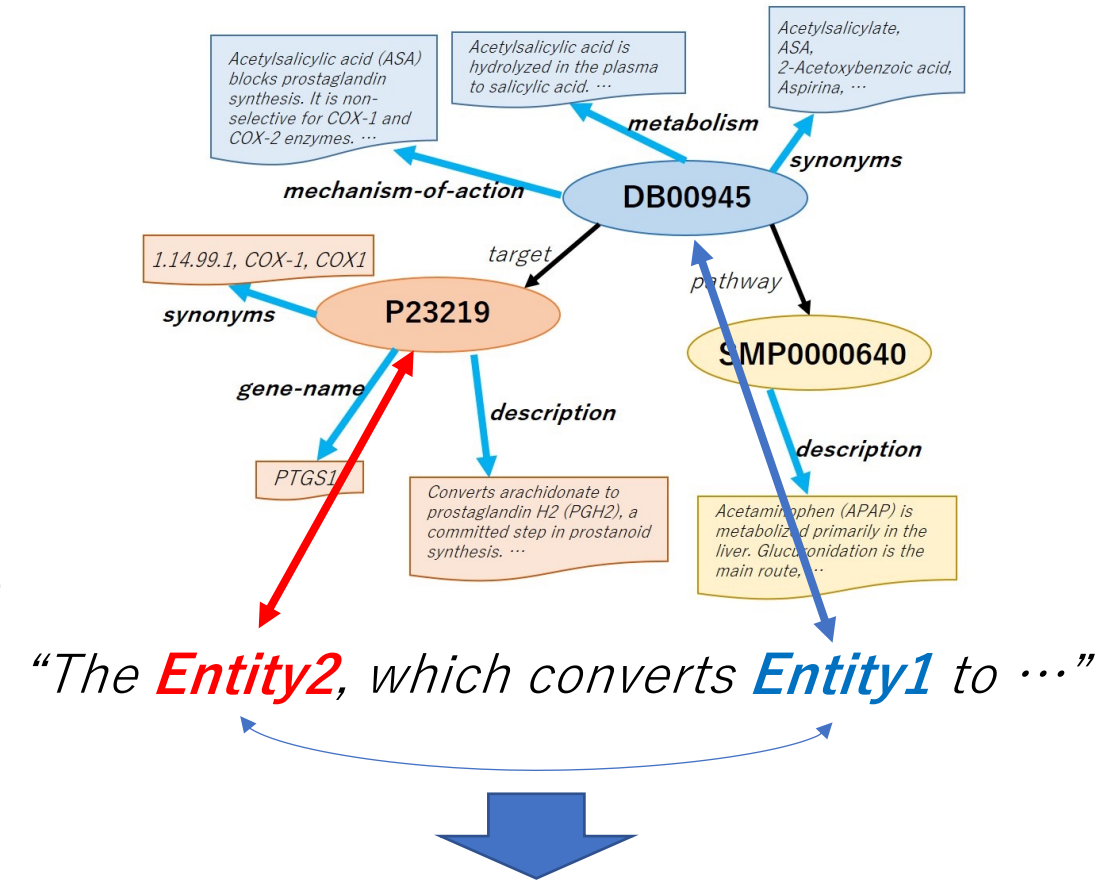
専門分野知識データベースには様々な意味も表現も異なる**ヘテロな**情報が含まれている

薬物に関する情報の例

- 種類
- **説明文（言語情報）**
- 名称・一般名
- **化学構造（構造情報）**
- 標的タンパク質 など

薬学分野における専門分野知識の情報抽出への利用

- 深層学習により様々な専門分野知識を統合し，情報抽出に利用
1. 言語・構造などそれぞれを扱う深層学習モデル
2. ヘテロな分野知識を用語に統合する深層学習モデル
3. 専門分野知識を情報抽出に利用する深層学習モデル
- 薬物関係抽出における抽出性能（F値）を83.70% から85.41%に向上 (世界最高性能)
  - 約7千文の学習データ+数百万の関係を持った概念グラフ
- 人の一致率を超える性能



Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Integrating heterogeneous knowledge graphs into drug-drug interaction extraction from the literature. Bioinformatics, Vol. 39, Issue 1, btac754, 2023.

今後の取り組み

- 文献からの抽出した情報と専門分野知識の連携
- 抽出した情報を利用したアプリケーション