講演資料

第2回スマート情報技術研究センターシンポジウム
第18回ジョイントCSセミナー

物質工学文献からの情報抽出のための コーパス構築とその評価

佐々木裕
Yutaka Sasaki

豊田工業大学　知能数理研究室
Computational Intelligence (COIN) Laboratory

References

- Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki, SC-CoMIcs: A Superconductivity Corpus for Materials Informatics, *12th Language Resources and Evaluation Conference (LREC-2020),* pp. 6753-6760, May, 2020.
- 山口京佑, 旭良司, 佐々木裕，文献抄録中の主題材料に着目した超伝導材料に関する情報抽出，言語処理学会第27回年次大会(NLP2021)，March, 2021.
- Kyosuke Yamaguchi, Ryoji Asahi, Yutaka Sasaki, Superconductivity information extraction from the literature: a new corpus and its evaluations, *Advanced Engineering Informatics*, Volume 54, ISSN 1474-0346, Elsevier, October 2022.

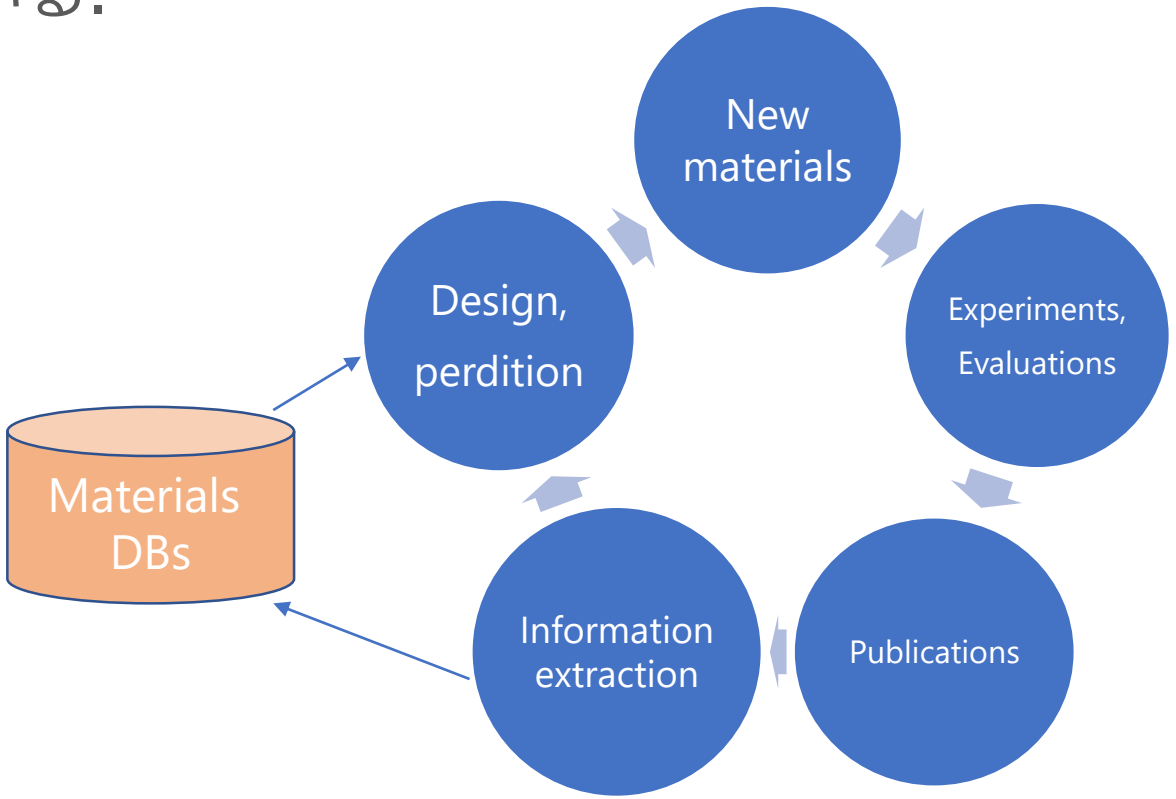
背景

マテリアルズ・インフォマティクス Materials Informatics (MI)

- 新材料の探索・発見の加速のために情報技術を物質工学分野に応用する.
- 近年，急速に発展し注目を集めている.

✓物質系データーベース

- Example: NIMS MatNavi[*1]
  ➡ 代表的な公開データベース
  ➡ 日々論文が出版



半自動的なデータベース構築が求められている

*1 https://mits.nims.go.jp/

背景｜情報抽出

非構造データである自由記述テキストから
所望の情報を構造化した形で自動抽出するタスク

- 固有表現抽出，関係抽出，イベント抽出

背景｜情報抽出

非構造データである文献テキストから
所望の情報を構造化した形で自動抽出するタスク

- **固有表現抽出**，関係抽出，イベント抽出

固有名や定量表現などを抽出するタスク

MgB_2 is a superconductor that have been applied to MRI .

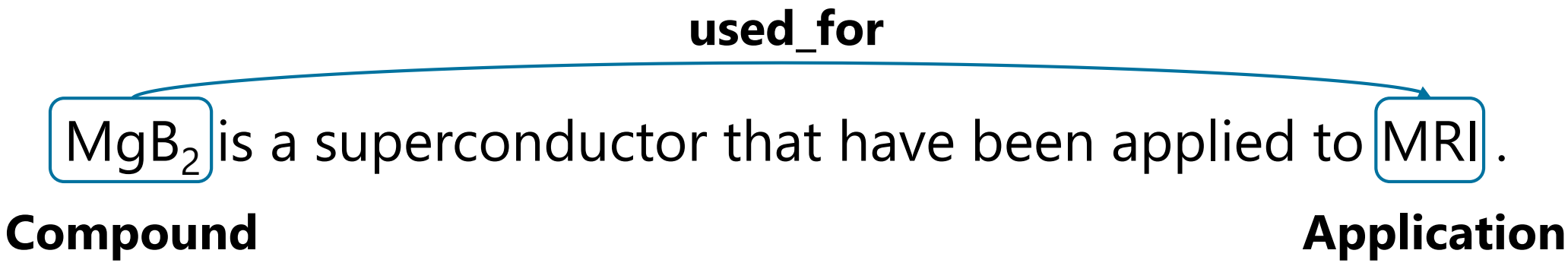
Compound **Application**

背景｜情報抽出

非構造データである文献テキストから
所望の情報を構造化した形で自動抽出するタスク

- 固有表現抽出，**関係抽出**，イベント抽出

固有表現間の関係を特定するタスク



背景｜情報抽出

非構造データである文献テキストから
所望の情報を構造化した形で自動抽出するタスク

- 固有表現抽出，関係抽出，**イベント抽出**

操作や動作などの事象を抽出するタスク

YBCO is synthesized by sintering a mixture of Y_2O_3, $BaCO_3$, and CuO.

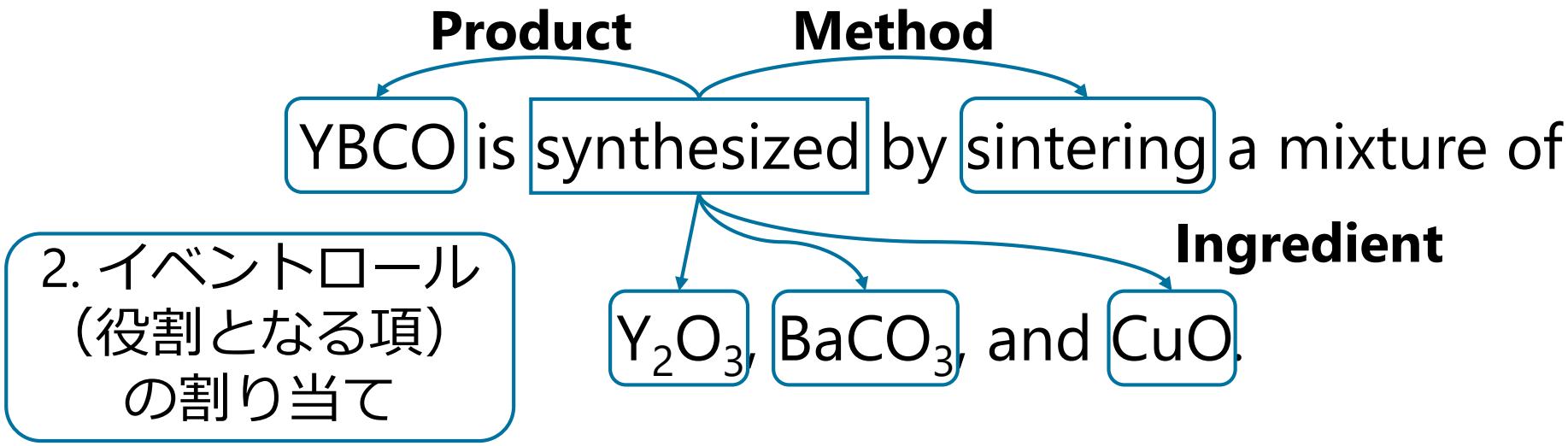
1. トリガーワード
（イベント発生）

背景｜情報抽出

非構造データである文献テキストから
所望の情報を構造化した形で自動抽出するタスク

- 固有表現抽出，関係抽出，**イベント抽出**

操作や動作などの事象を抽出するタスク



背景｜超伝導材料と情報抽出

超伝導材料

- ✓医療機器やリニア新幹線といった幅広い分野で応用されている
- ✓より高い転移温度で超伝導を発現する高温超伝導体の開発が目指されている

超伝導材料に関する情報抽出

- 問題：利用可能なタグ付きコーパスがほとんど存在しない（2017年時点）
- ※ タグ付きコーパス：
  - テキストに抽出したい情報を注釈付けしたデータセット（教師データ）

背景｜超伝導材料と情報抽出

超伝導材料

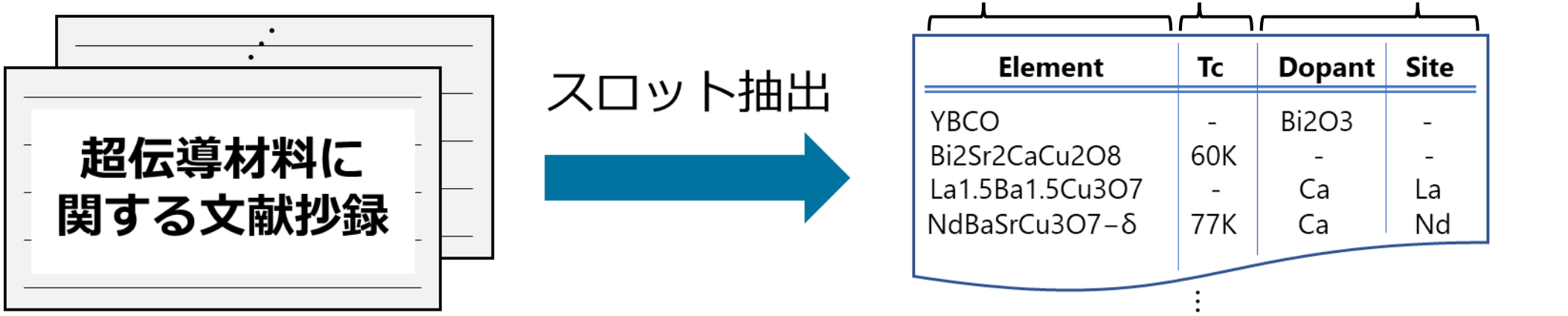
- ✓医療機器やリニア新幹線といった幅広い分野で応用されている
- ✓より高い転移温度で超伝導を発現する高温超伝導体の開発が目指されている

超伝導材料に関する情報抽出

- 問題：利用可能なタグ付きコーパスがほとんど存在しない（2017年時点）
- ※ タグ付きコーパス：
  - テキストに抽出したい情報を注釈付けしたデータセット（教師データ）

目的

✓文献抄録から超伝導材料に関する情報を構造化した形で抽出



材料組成・転移温度・ドーピング情報をまとめて抽出
➡ 高度なデータ分析への活用が可能

提案手法

超伝導材料に関する情報抽出システムを提案

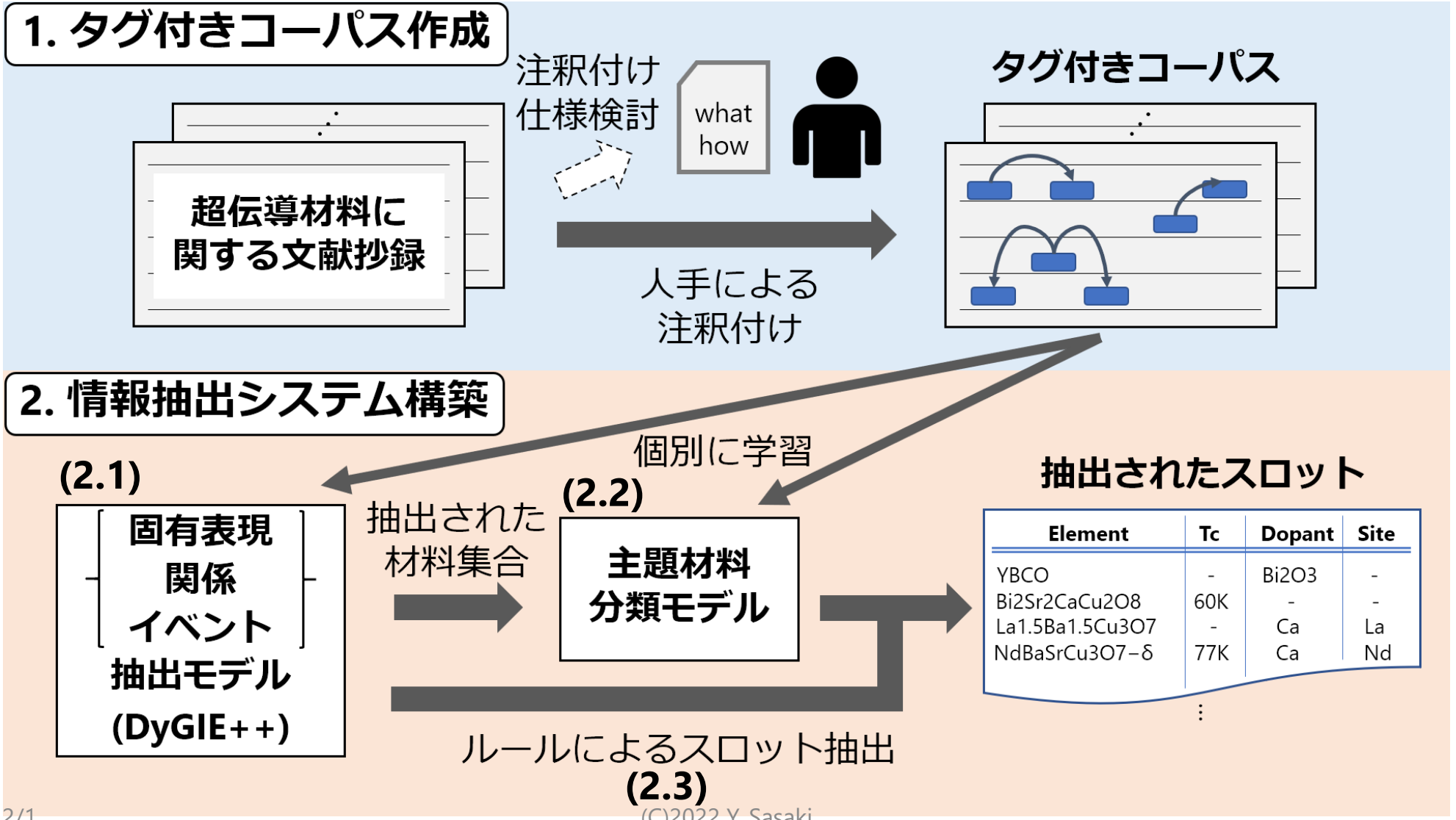
1. タグ付きコーパスの作成
   - システム構築にはデータリソースとなるタグ付きコーパスが必要
   - 我々の目的に合ったコーパスを作成：抽出したい情報を独自に定義・注釈付け
2. スロット抽出（抽出対象：材料組成・転移温度・ドーピング情報）
   - 材料組成に対する転移温度・ドーピング情報の紐づけにより実現

   問題：DyGIE++における関係抽出では文内のみを対象 ➡ 文を跨いだ紐づけ✕

   ✓抄録中で**話題として取り上げられている材料（主題材料）**を特定

   ➡ 文内で紐づけられなかった転移温度・ドーピング情報を
   主題材料に暗黙的に紐づけすることで情報の取りこぼしを低減する狙い

提案手法（超伝導材料に関する情報抽出システム）



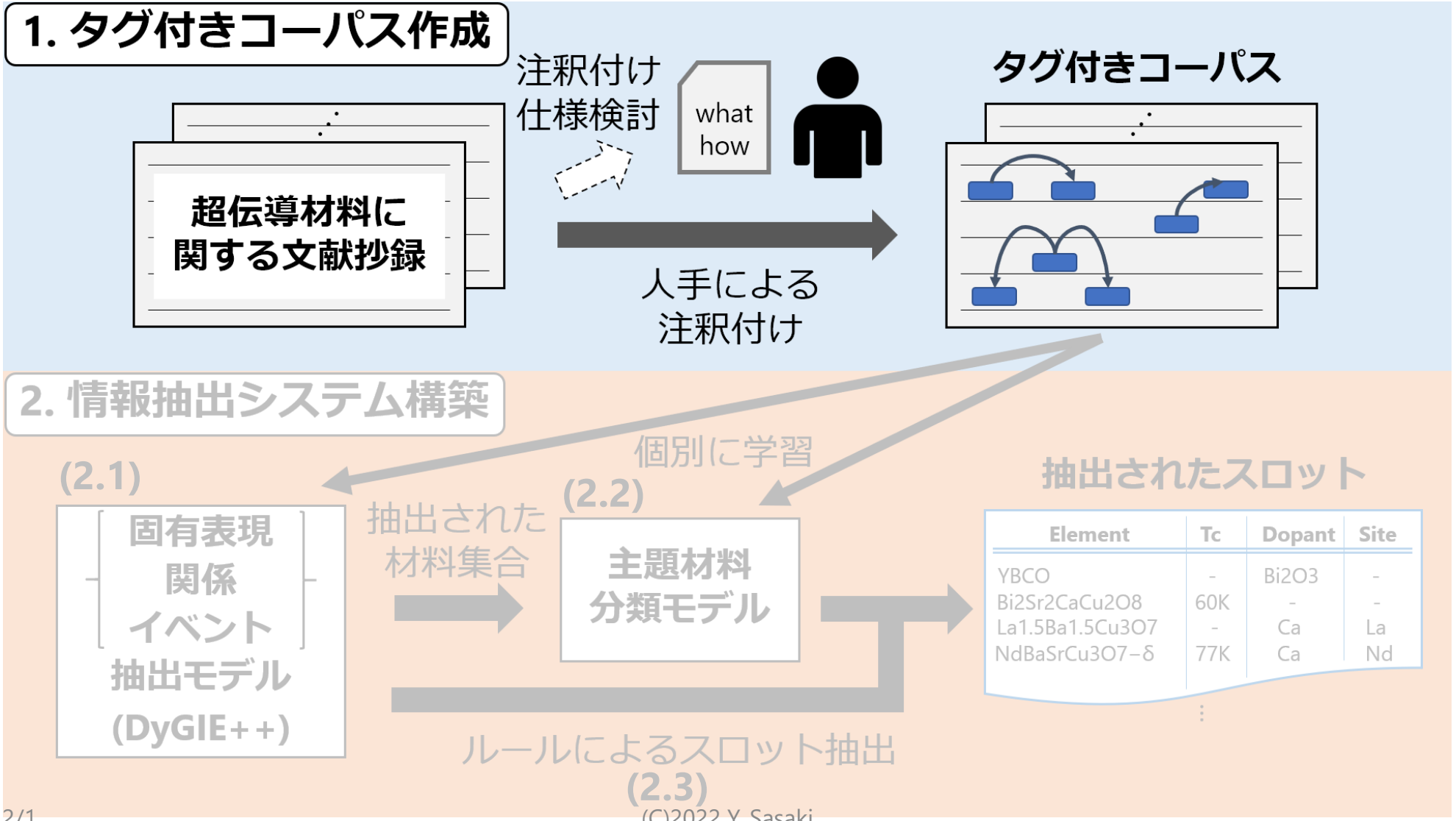
コーパス構築
Corpus construction

SC-CoMIcs

SuperConductivity Corpus for Materials Informatics

(doi: 10.17632/xc9fjz2p3h.2)

提案手法 (1)｜タグ付きコーパスの作成



定義した注釈付け仕様の概要

- 超伝導材料に関する**基本情報**（固有表現クラス）
  - 一般材料に共通する６つ，超伝導に関する１つ
- 文献抄録中の**主題材料**（固有表現クラス）
- 超伝導材料に固有の情報
  - **転移温度**（関係クラス）
  - **ドーピング情報**（イベントクラス）
- 文内における**材料組成・転移温度・ドーピング情報の紐づけ**
  - 材料組成と転移温度・ドーピング情報を紐づけ（関係クラス）
  - 転移温度とドーピング情報の紐づけ（関係クラス）

仕様定義｜超伝導材料に関する基本情報

７つの固有表現クラスを定義

	クラス名	定義	例
一般材料に共通	Characterization	**分析手法**に関する用語	X-ray diffraction, SEM
	Process	**合成プロセス**に関する用語	sol-gel, calcination
	Property	**材料特性・物性理論**に関する用語	electrical, magnetic fields
	Material	**結晶構造**に関する用語	tetragonal, bulk, film
	Element	**元素名**と**化合物名**	Ti, oxygen, $YBa_2Cu_3O_7$
	Value	**定量表現**（単位含む）	45%, 95K, 0.08
	SC	**超伝導**に関する用語	superconductivity, Tc

仕様定義｜主題材料の定義（固有表現）

文献抄録において話題として取り上げられている材料

- Elementクラスの固有表現の内，
  文献中の主題であるものを新たなMainクラスで上書きする形で定義

（抄録の例）

In this work, we report on the effect of Y3+ doping on structural,

mechanical and electrical properties of Bi-2202 phase.

⋮

主題材料
（Mainクラス）

仕様定義｜転移温度（関係）

常伝導から超伝導，もしくは超伝導から常伝導に相転移する際の温度

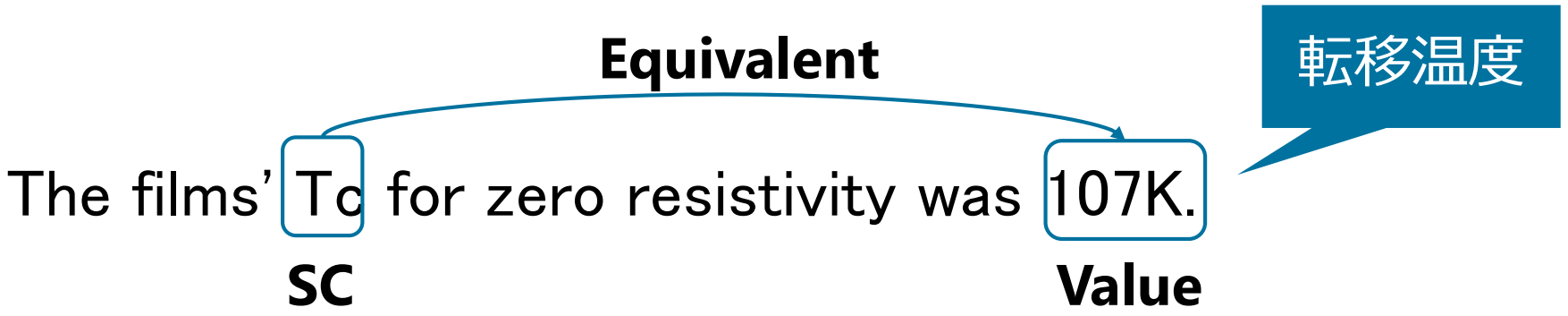
The films' Tc for zero resistivity was 107K.

SC **Value**

仕様定義｜転移温度（関係）

常伝導から超伝導，もしくは超伝導から常伝導に相転移する際の温度

✓固有表現クラスSCからValueに対して関係クラスEquivalentを付与



仕様定義｜ドーピング情報（イベント）

ドーピング：材料の物質特性を変化させるために不純物を添加する操作

✓不純物はドーパント，ドーパントの受け入れ先はサイトと呼ばれる

Ni doping into the CuO_2 plane

仕様定義｜ドーピング情報（イベント）

ドーピング：材料の物質特性を変化させるために不純物を添加する操作

✓不純物はドーパント，ドーパントの受け入れ先はサイトと呼ばれる

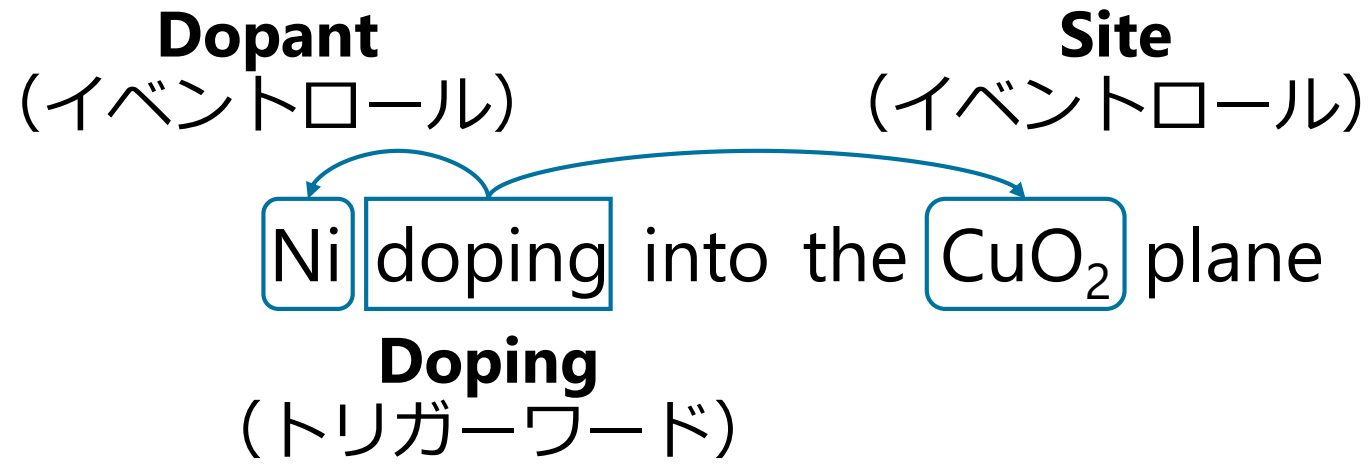
Ni **doping** into the CuO_2 plane

Doping
（トリガーワード）

仕様定義｜ドーピング情報（イベント）

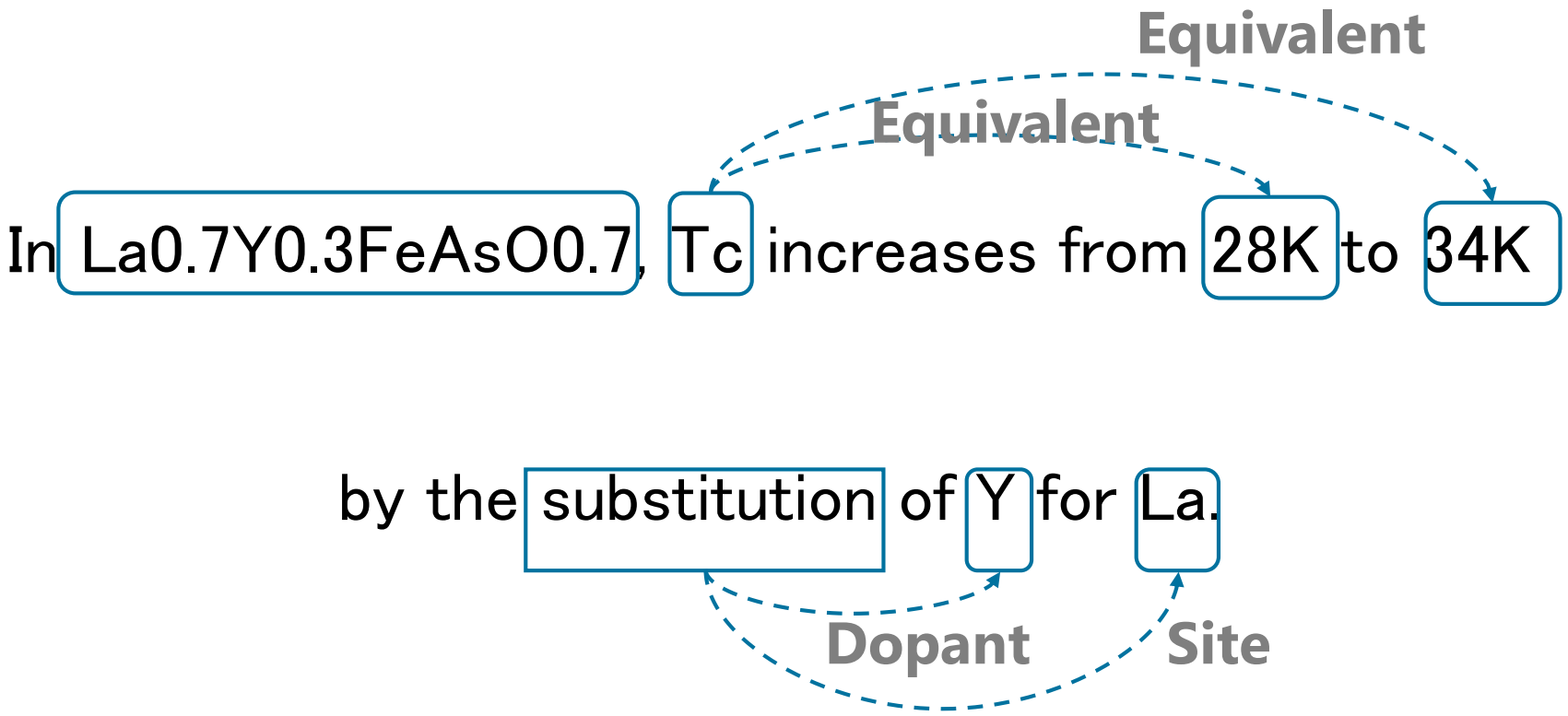
ドーピング：材料の物質特性を変化させるために不純物を添加する操作

✓不純物はドーパント，ドーパントの受け入れ先はサイトと呼ばれる



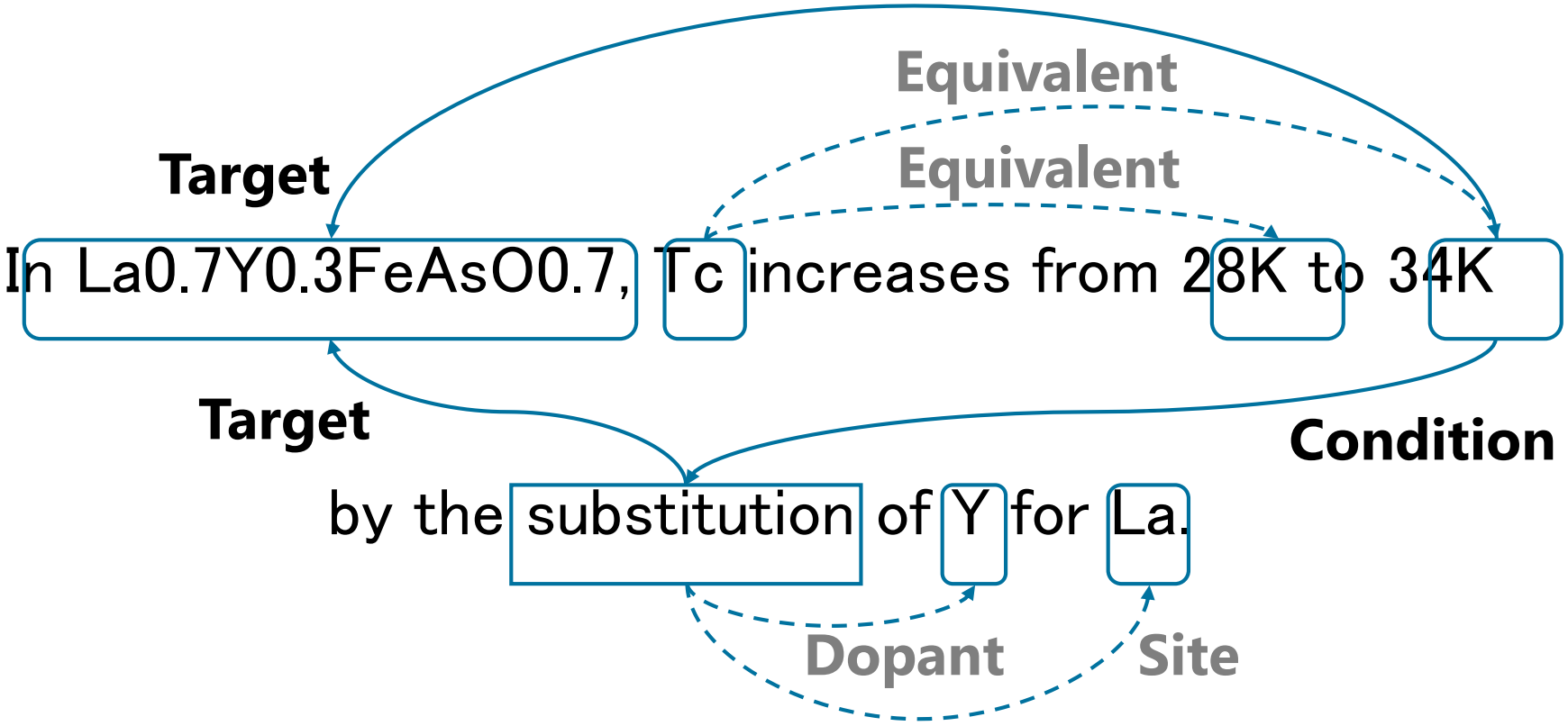
仕様定義｜スロット抽出対象の紐づけ（関係）

スロット抽出対象：材料組成・転移温度・ドーピング情報



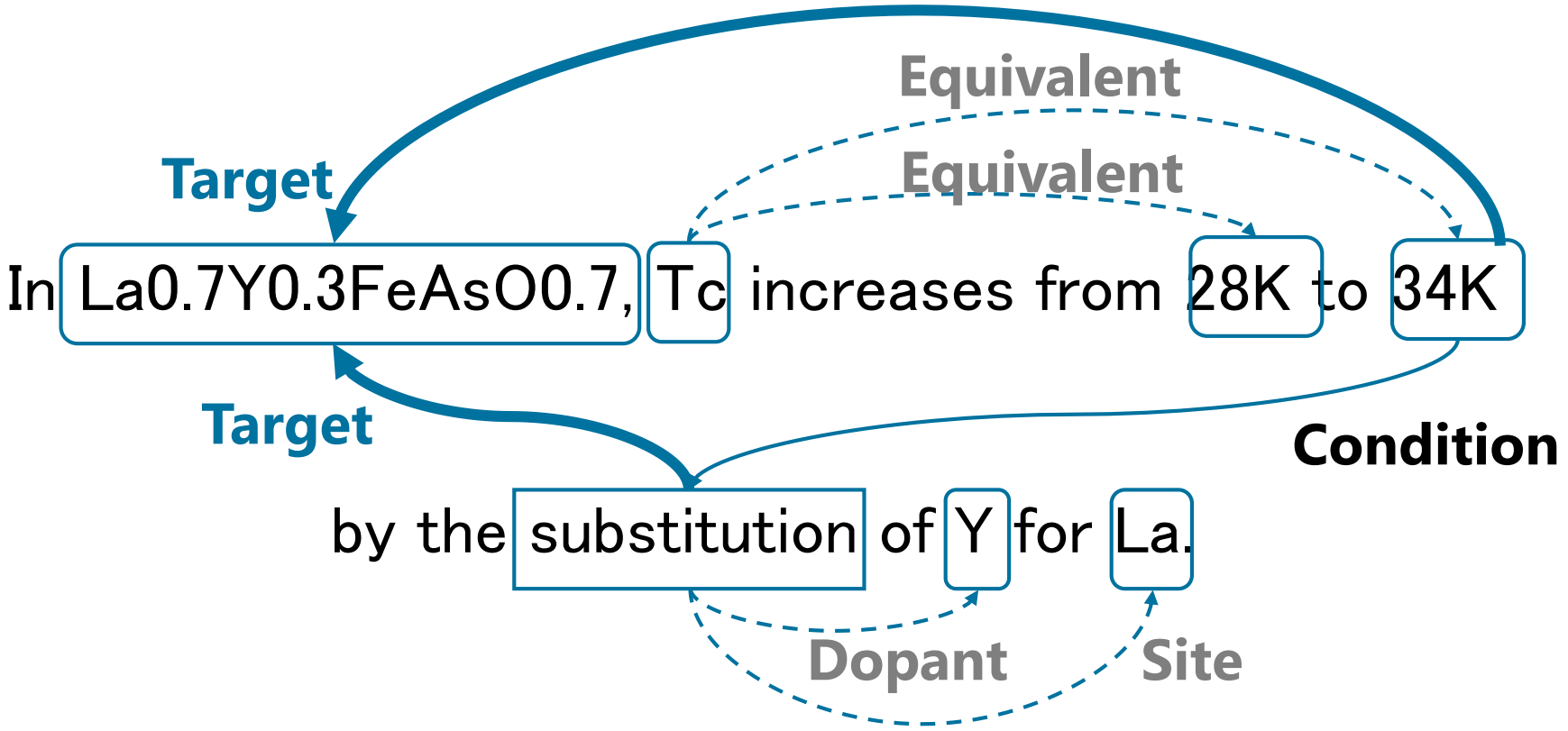
仕様定義｜スロット抽出対象の紐づけ（関係）

スロット抽出対象：材料組成・転移温度・ドーピング情報



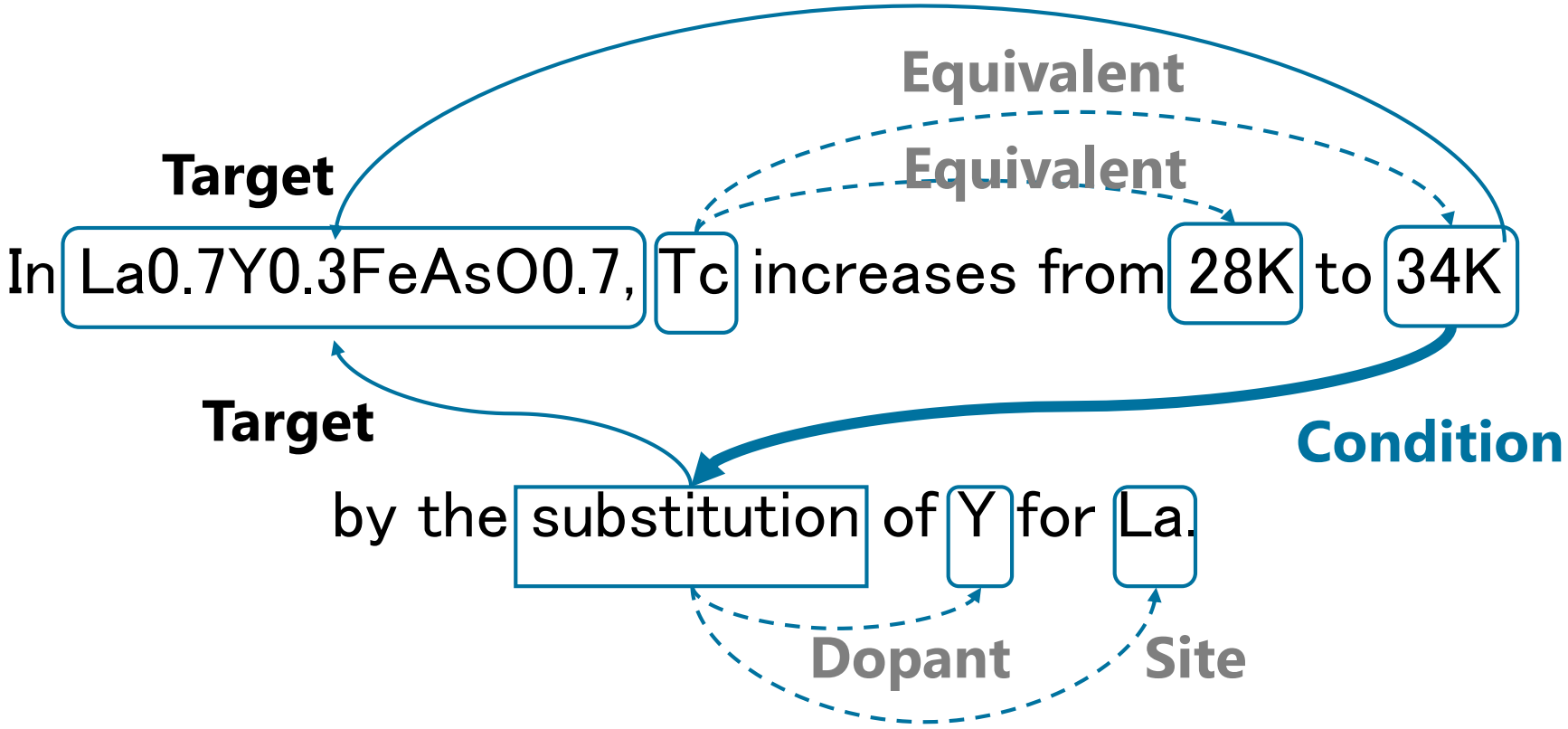
仕様定義｜スロット抽出対象の紐づけ（関係）

材料組成に対する転移温度・ドーピング情報の紐づけ（文内のみ対象）



仕様定義｜スロット抽出対象の紐づけ（関係）

転移温度とドーピング情報の紐づけ（文内のみ対象）



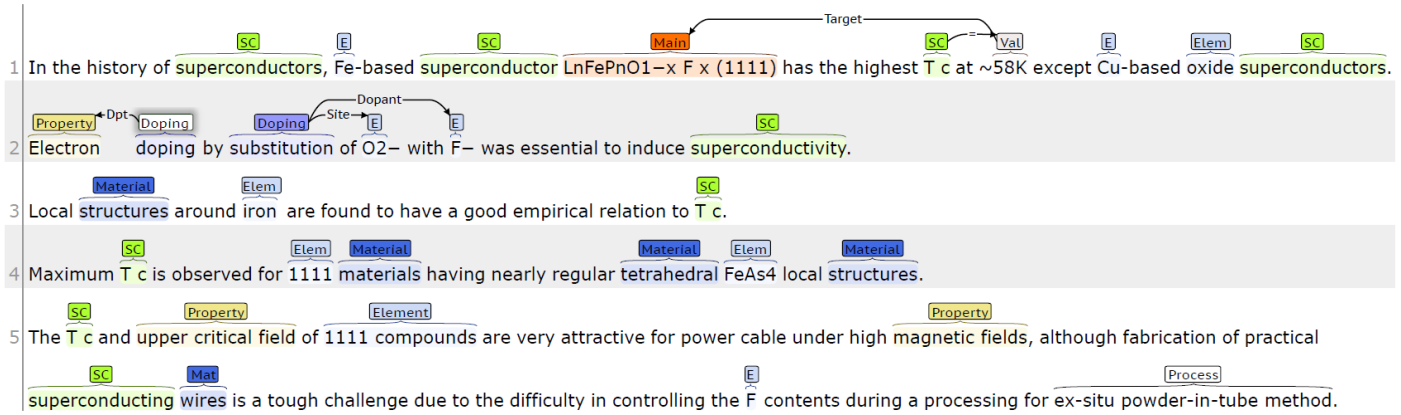
抄録収集・人手による注釈付け

抄録収集（２回に分けて実施）

- Science Direct のWeb APIを利用
- １回目：200件 ➡ 仕様検討用
- ２回目：800件 ➡ 注釈付け用

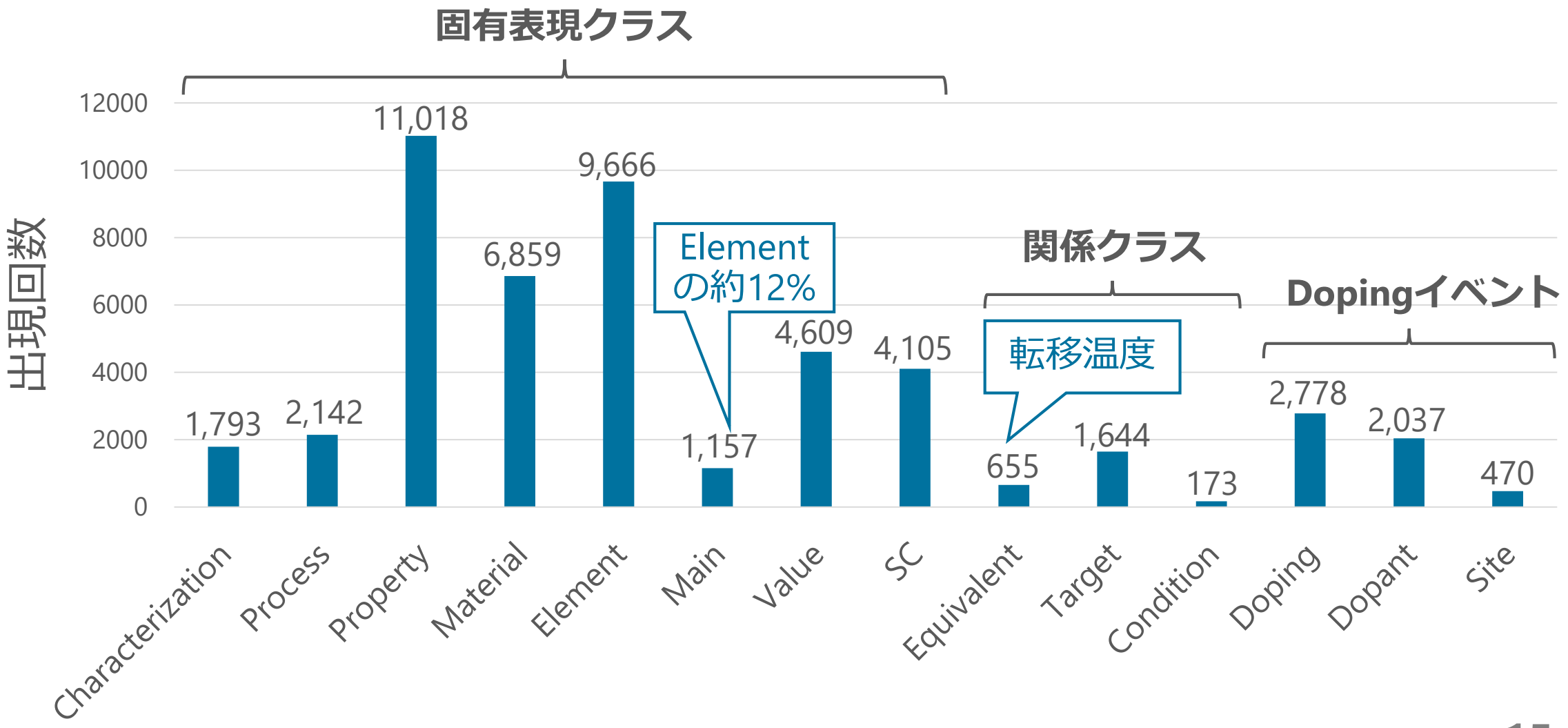
人手による注釈付け

- Value・Main以外の固有表現クラス
- Valueクラス，関係クラス，Dopingイベント
- Mainクラス



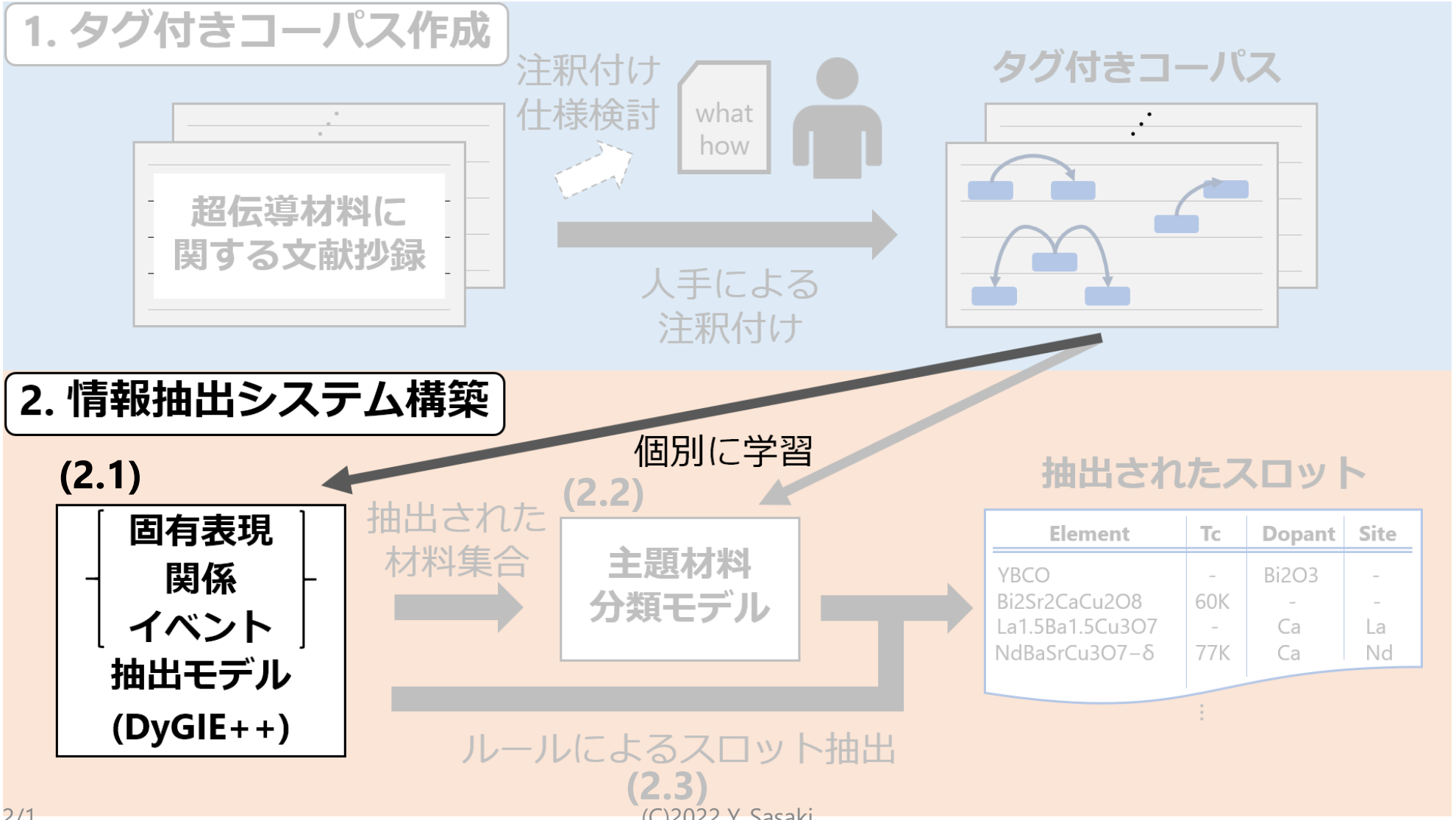
抄録１件に対して注釈付けした結果

実験 (1)｜タグ付きコーパスの解析



評価実験
Evaluation Experiments

提案手法 (2.1)｜固有表現・関係・イベント抽出

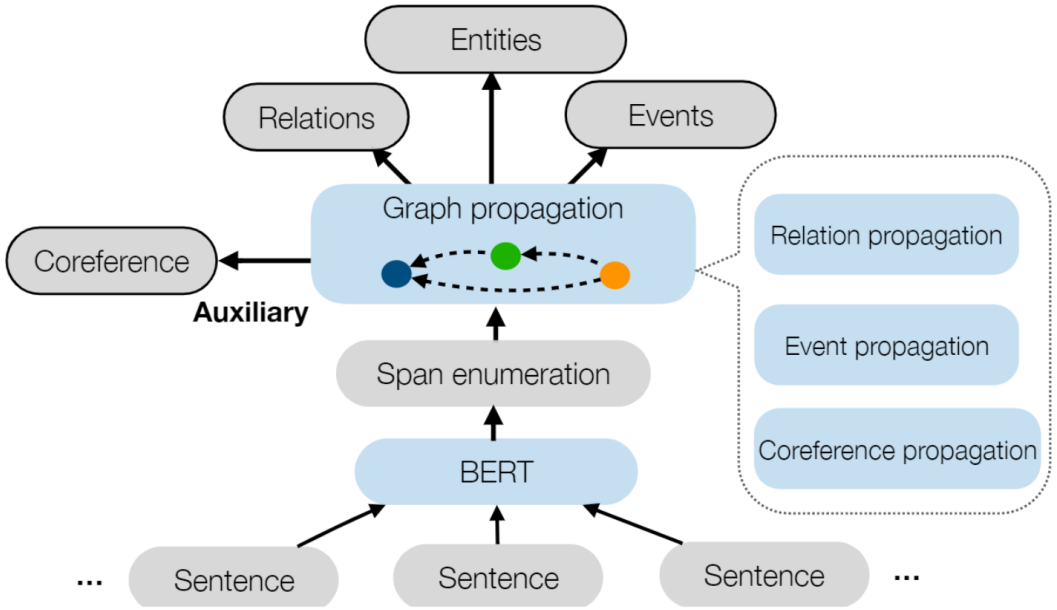


関連研究｜DyGIE++ [Wadden+ 2019]

ニューラルネットワークを用いた
固有表現・関係・イベントの同時抽出手法

- それぞれのタスクを解く上で重要となる情報を相互にグラフ伝播する機構により各タスクで高い抽出精度を達成

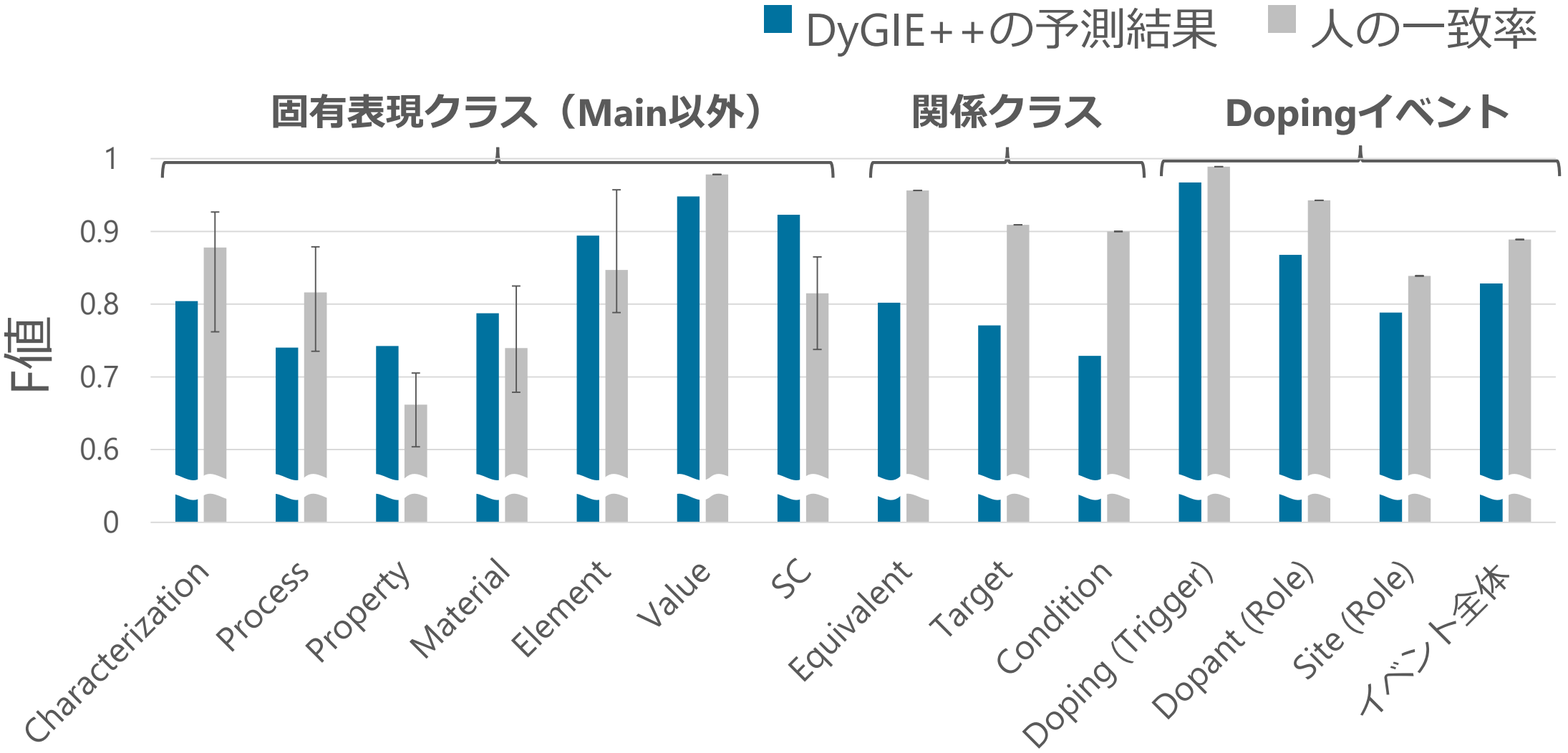
✓関係抽出は**文内のみ**を対象



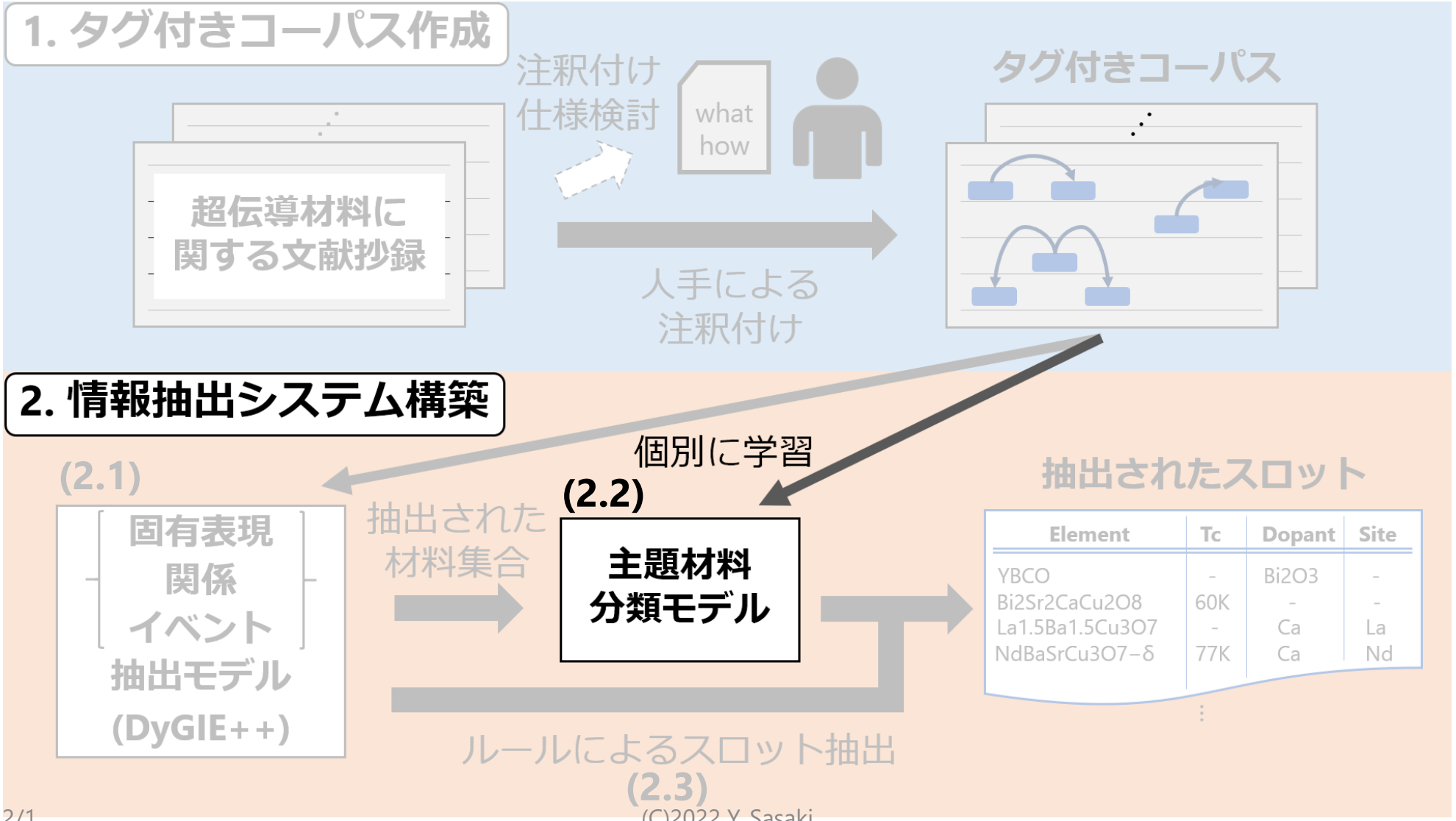
DyGIE++のモデル概要

David Wadden et al. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789.

実験 (2.1)｜固有表現・関係・イベントの抽出精度



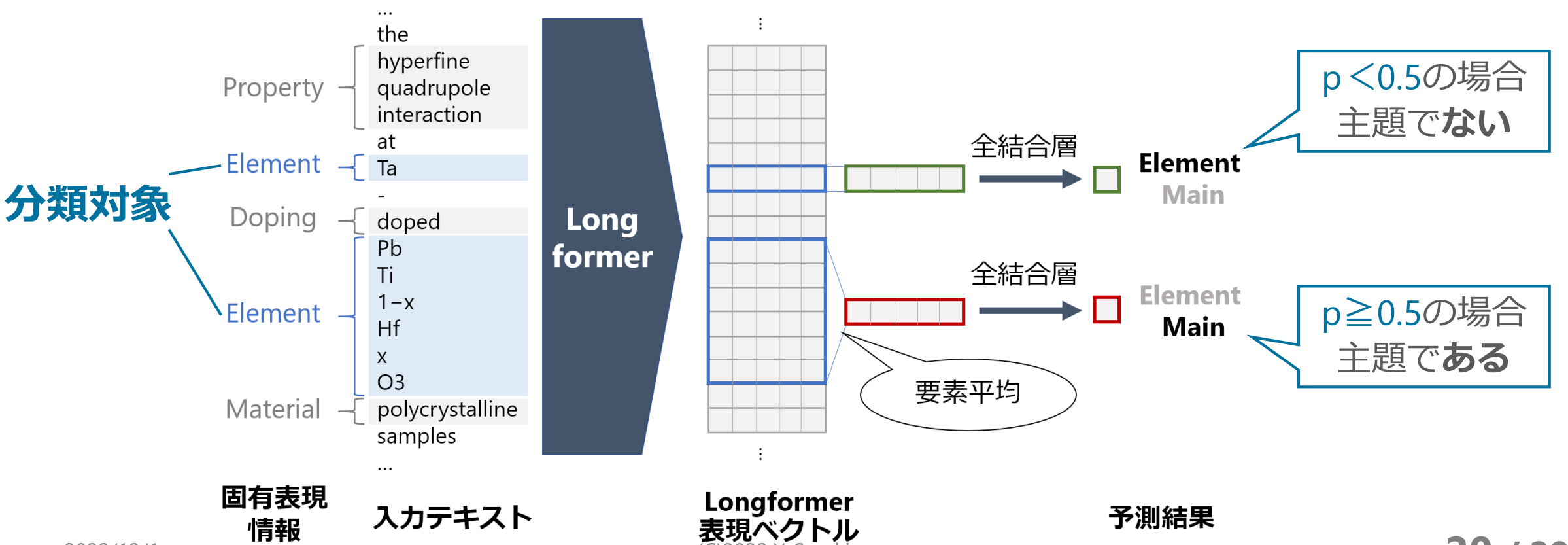
提案手法 (2.2)｜主題材料分類



主題材料分類モデル

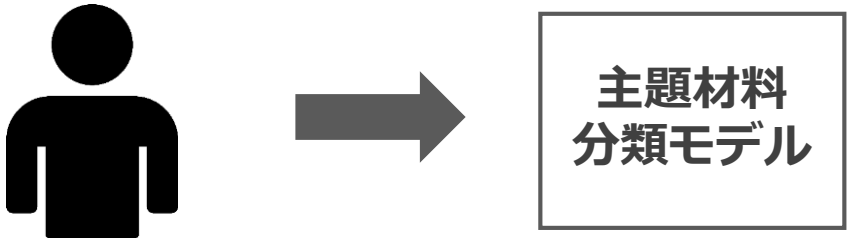
Elementクラスの固有表現を対象に「主題で “ある” or “ない”」で二値分類

- Longformer：長い文脈を捉えることに優れたニューラルネットワーク



実験 (2.2)｜主題材料の分類精度

✓主題材料分類モデルへの入力とするElement情報を変えて検証



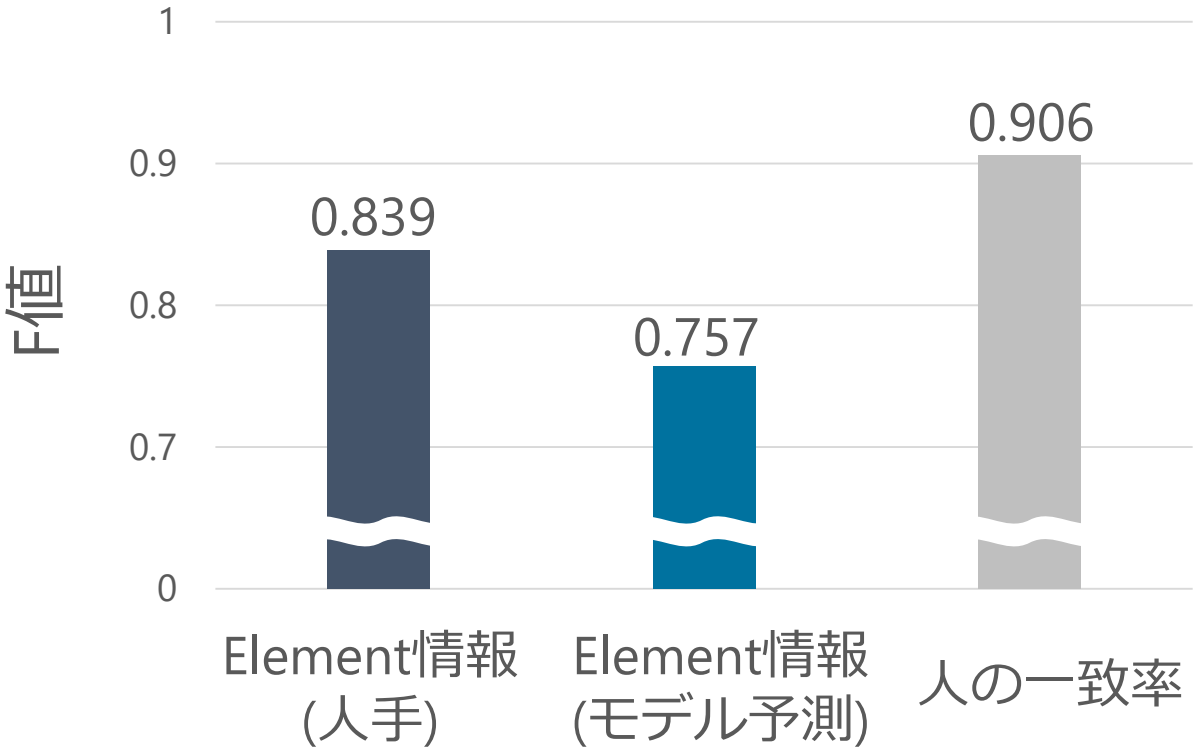
主題材料分類モデル

Element情報：
人手による注釈結果

固有表現抽出モデル (DyGIE++)

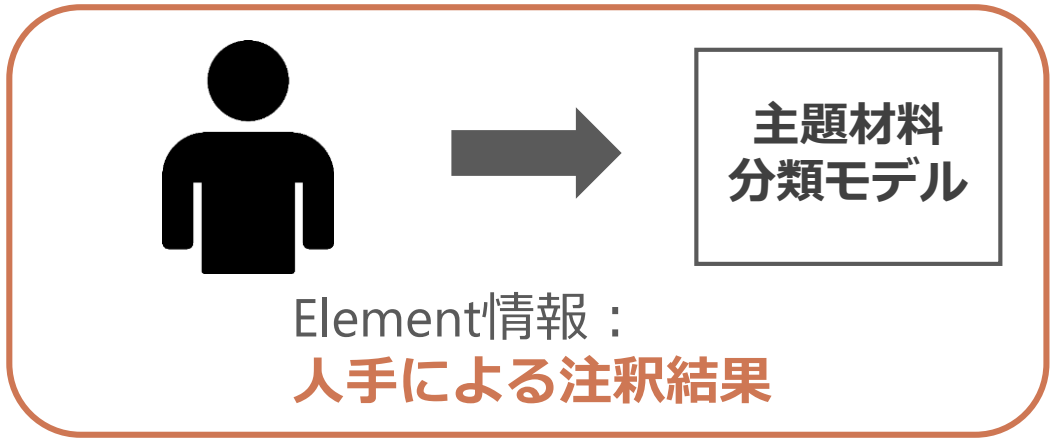
主題材料分類モデル

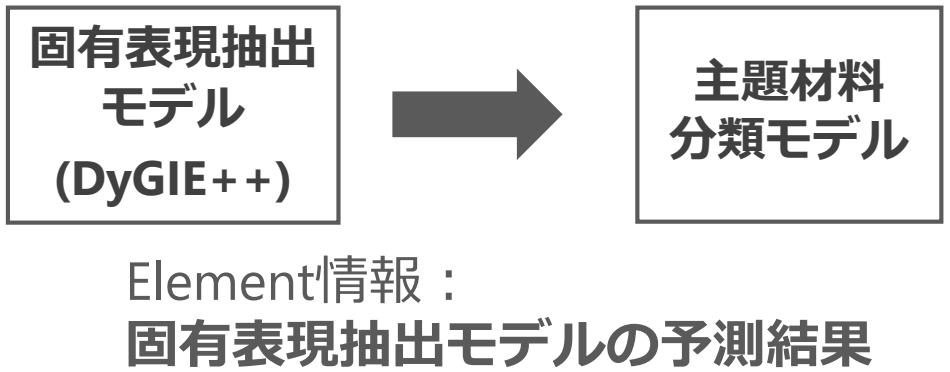
Element情報：
固有表現抽出モデルの予測結果

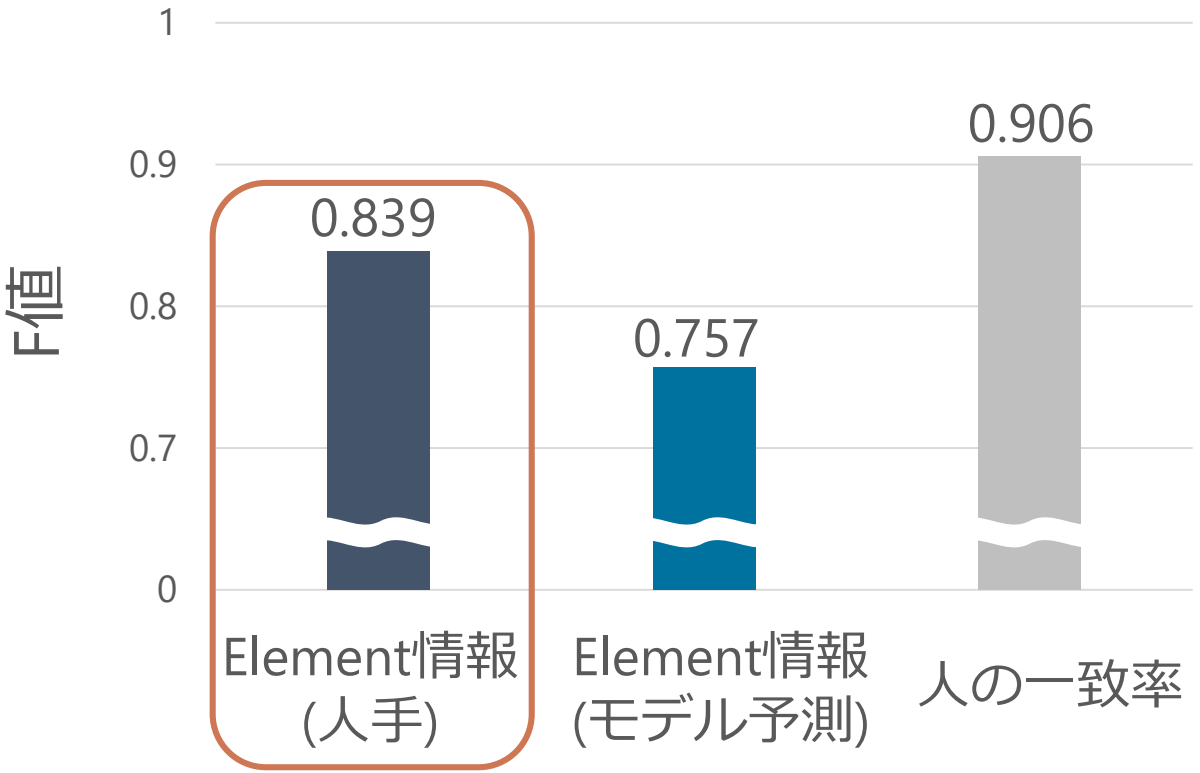


実験 (2.2)｜主題材料の分類精度

✓主題材料分類モデルへの入力とするElement情報を変えて検証

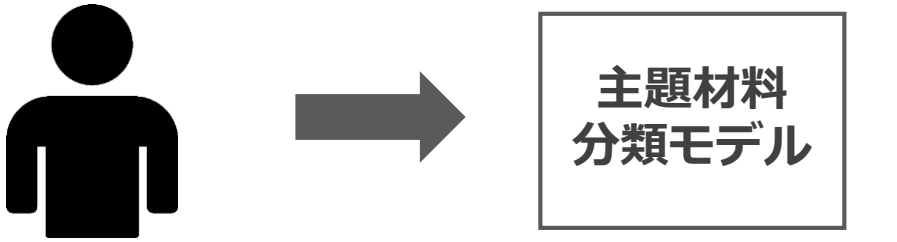






実験 (2.2)｜主題材料の分類精度

✓主題材料分類モデルへの入力とするElement情報を変えて検証



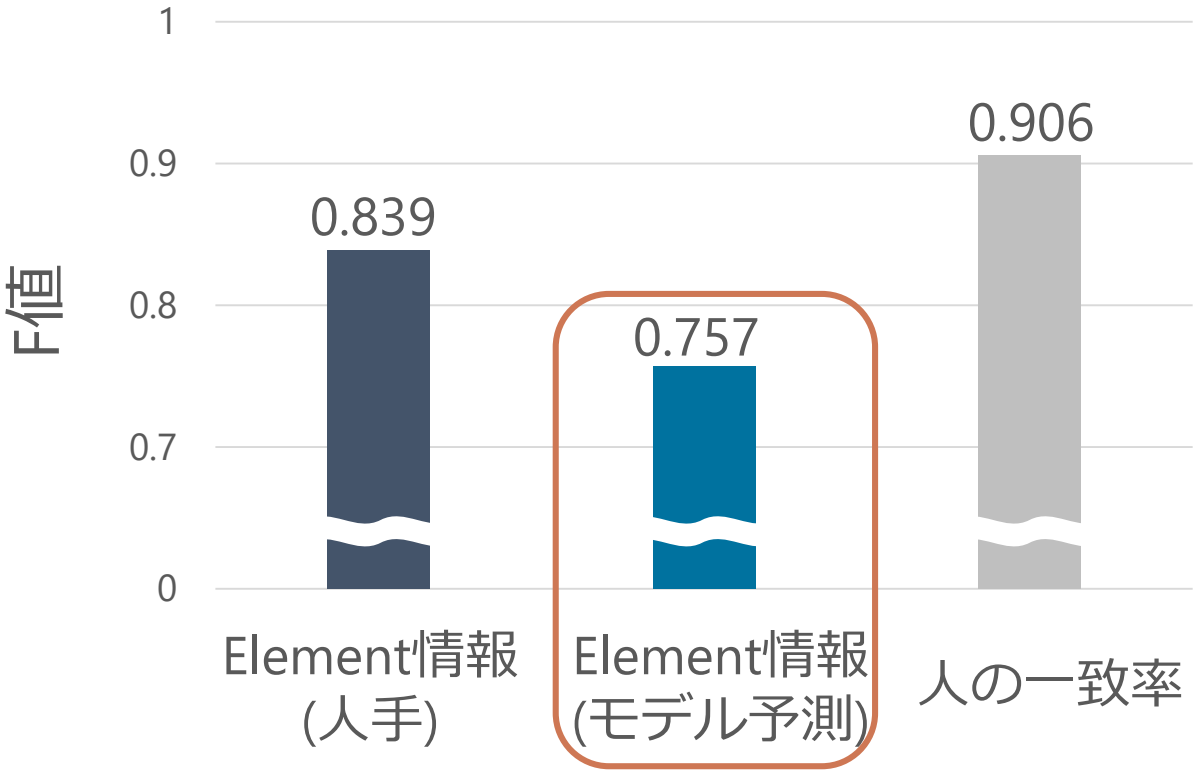
主題材料分類モデル

Element情報：
人手による注釈結果

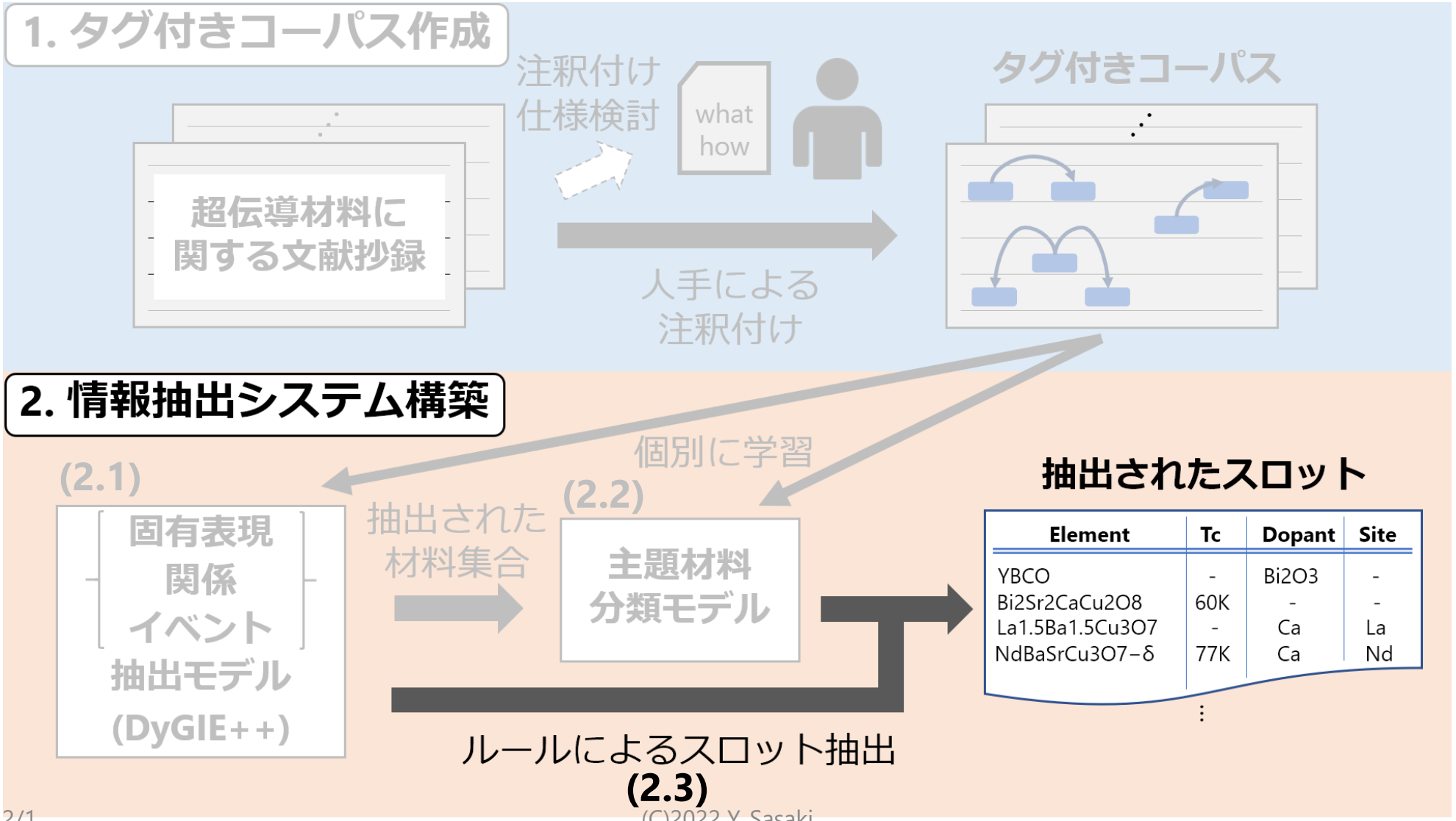
固有表現抽出モデル (DyGIE++)

主題材料分類モデル

Element情報：
固有表現抽出モデルの予測結果



提案手法 (2.3)｜ルールによるスロット抽出



材料組成への紐づけに関するルール

- 文内で関係付けされる場合

1文目 In this work, we report on the effect of Y3+ doping on structural, mechanical and electrical properties of Bi-2202 phase.

Dopant

Target

文内の関係のみを対象

材料組成への紐づけに関するルール

- 文内で関係付けされる場合

文内の関係のみを対象

Dopant

1文目 In this work, we report on the effect of Y3+ doping on structural, mechanical and electrical properties of Bi-2202 phase.

Target

Target関係に従って
スロット抽出

Element	Tc	Dopant	Site
Bi-2202 phase	-	Y3+	-

材料組成への紐づけに関するルール

- 文内で関係付けされない場合

文内の関係のみを対象

Dopant

1文目 In this work, we report on the effect of Y3+ doping on structural,

Target

mechanical and electrical properties of Bi-2202 phase.

⋮

9文目 The higher onset transition temperature is obtained for x =0.025

Equivalent

and is about 93.62K.

（転移温度）

材料組成への紐づけに関するルール

- 文内で関係付けされない場合

文内の関係のみを対象

Dopant

1文目 In this work, we report on the effect of Y3+ doping on structural,

Target

mechanical and electrical properties of Bi-2202 phase.

主題材料（Mainクラス）

⋮

暗黙的な紐づけ

9文目 The higher onset transition temperature is obtained for x =0.025

Equivalent

and is about 93.62K.

（転移温度）

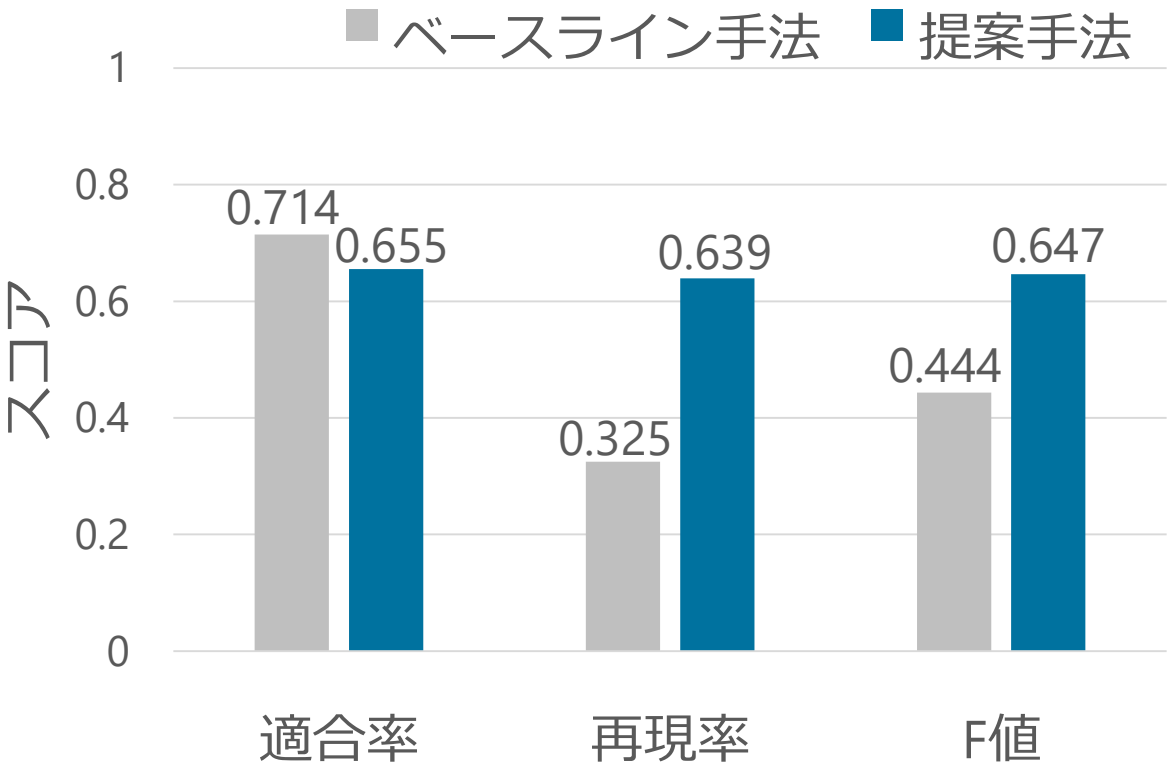
Element	Tc	Dopant	Site
Bi-2202 phase	93.62K	–	–

実験 (2.3)｜ルールによるスロット抽出性能評価

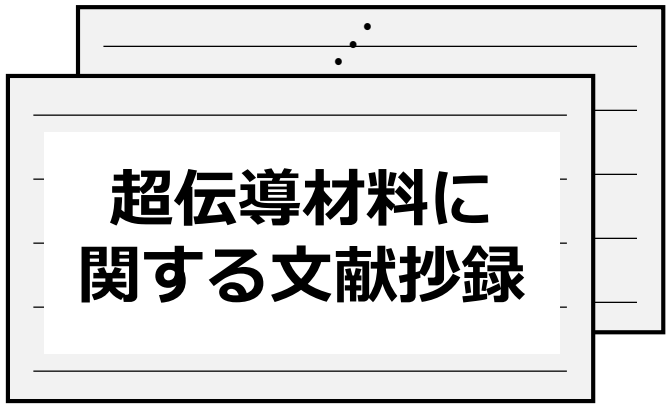
ルールによる主題材料への紐づけ精度
➡ F値：97.3% (ルールを適用して得られたスロットを正解として評価)

文内のみ（ベースライン手法）
- Targetによる文内での関係抽出の結果のみを用いて材料組成と紐づけ

文内＋文外（提案手法）
- 主題材料への暗黙的な紐づけも行う

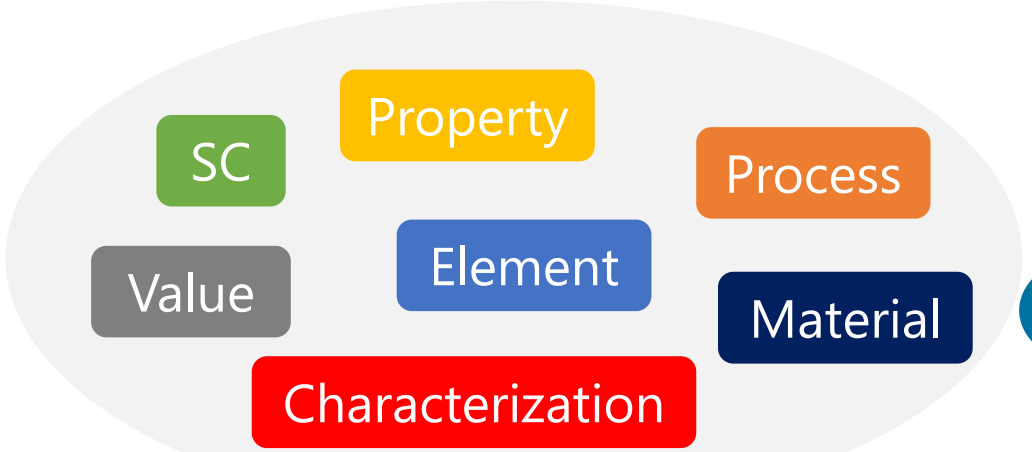


提案システムにより抽出されたデータの活用



提案システム
による情報抽出

抽出されたデータ



Element	Tc	Dopant	Site
YBCO	-	Bi2O3	-
Bi2Sr2CaCu2O8	60K	-	-
La1.5Ba1.5Cu3O7	-	Ca	La
NdBaSrCu3O7−δ	77K	Ca	Nd

様々な用途に活用可能
- ツール開発
- データ分析 etc.

抽出データ活用例①｜類似用語検索ツール

✓追加抄録8,898件を用意

- 提案システムで固有表現を抽出
  ➡クラス毎の用語リストを作成
- 元の1,000件と併せてword2vec(単語のベクトル化手法)を学習

（ベクトルの内積を取って類似度を計算）

検索対象とする
用語クラスを指定可能

（入力画面）

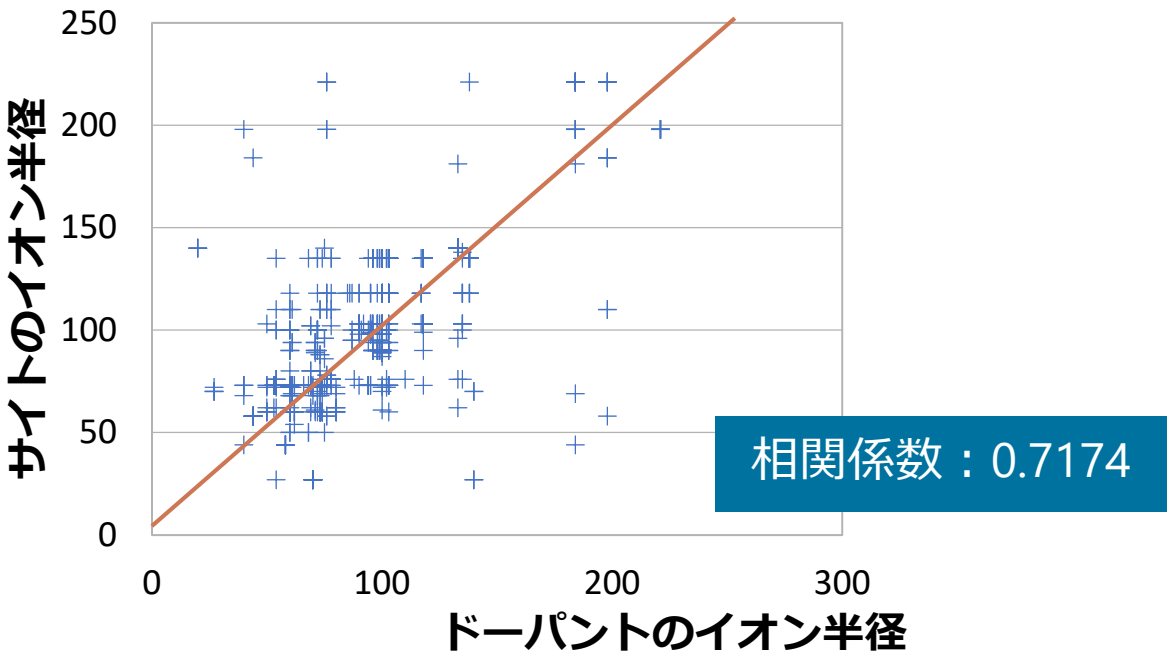
（検索結果）

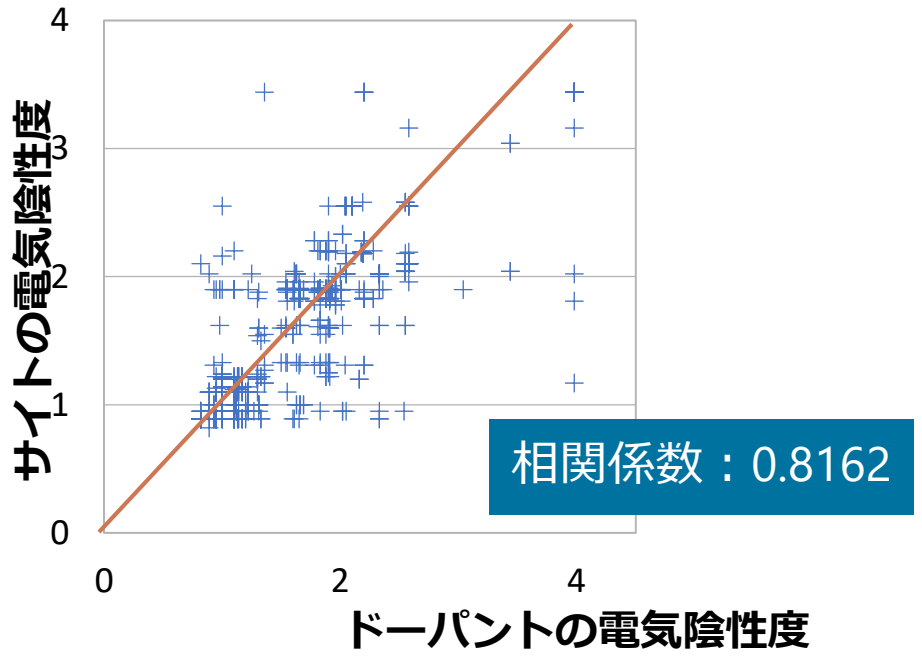
rank	word	score
1	surface layers	0.35306
2	local structure	0.35212
3	structural distortion	0.35125
4	epitaxial strain	0.33729
5	filamentary	0.33685
6	sub-lattice	0.33661
7	stripe-type	0.33165
8	blocking layer	0.32527
9	structure distortion	0.32527
10	lattice structure	0.32331

抽出データ活用例②｜ドーピング情報の分析

Hume-Rotheryの法則： ２つの金属 (化合物) の固溶度に関する法則

- イオン半径・電気陰性度 ➡ ２つの差が小さいほど固溶度が大きくなる





どちらも0.7以上で強い相関を示した

まとめと今後の課題

まとめ

- 目的：文献抄録から超伝導材料情報を構造化した形で抽出
- 提案：抄録中の主題材料を活用した情報抽出システム
- 結果：最終的なスロット抽出精度は64.7%

✓提案システムを用いて抽出したデータは様々な用途に活用可能

今後の課題

- スロット項目の拡張
- 文書レベルでのシステム構築