第1週B4ゼミ

オントロジー構築

論文紹介

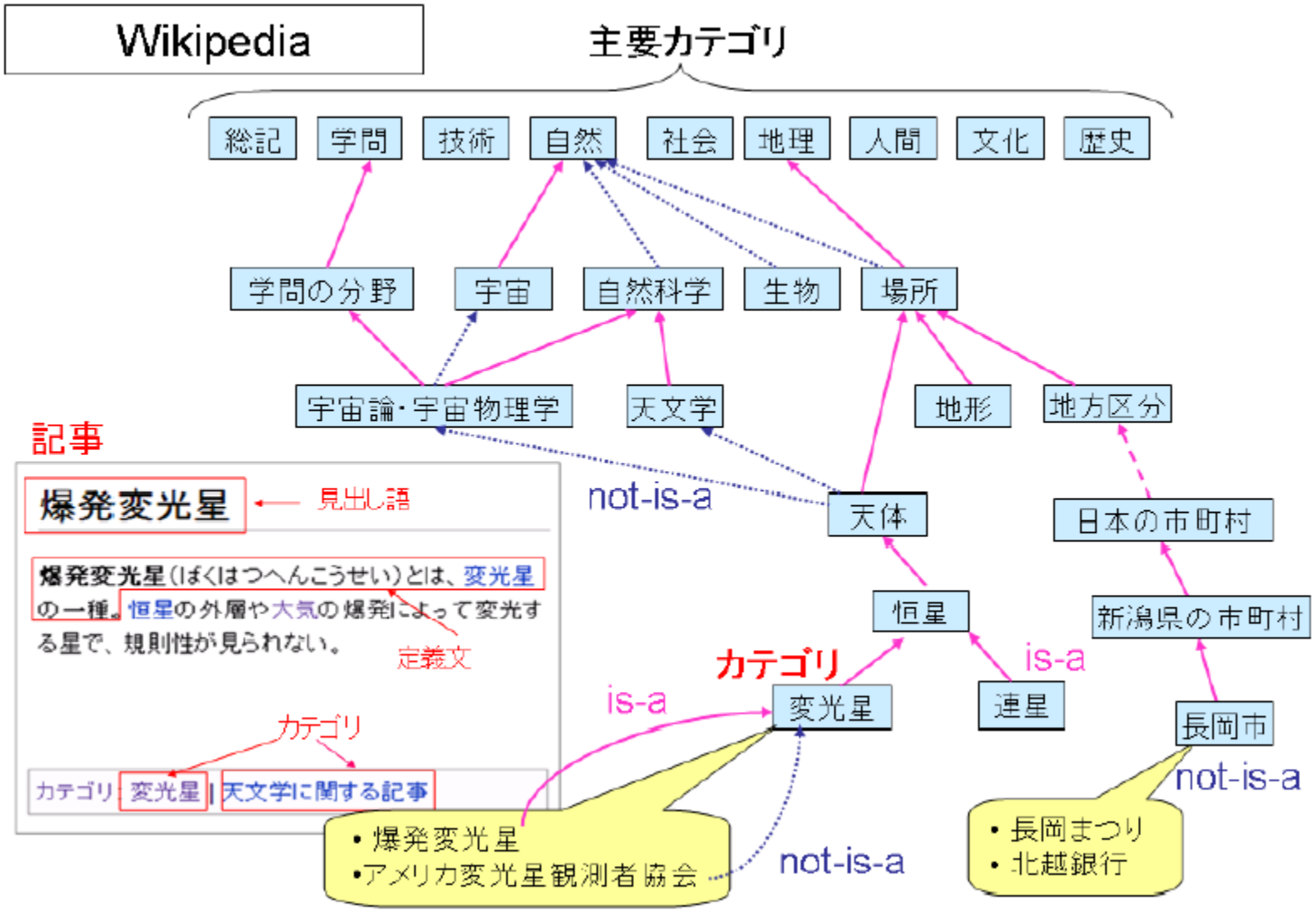
出典

- 柴木 優美　(2011.3)
- Wikipediaからの大規模な汎用オントロジー構築
- 長岡技術科学大学修士論文 (未公刊)

概要

- Wikipediaのis-a関係のリンクを判定
- 「人」、「組織」、「施設」、「地名」、「地形」、「具体物」、「創作物」、「動植物」、「イベント」
- 上記9 種類の意味属性を最上位カテゴリとしたis-a 関係のオントロジーの構築

Wikipedia


Wikipedia
主要カテゴリ
総記
学問
技術
自然
社会
地理
人間
文化
歴史
学問の分野
宇宙
自然科学
生物
場所
宇宙論・宇宙物理学
天文学
地形
地方区分
記事
爆発変光星
見出し語
爆発変光星（ばくはつへんこうせい）とは、変光星の一種。恒星の外層や大気の爆発によって変光する星で、規則性が見られない。
定義文
カテゴリ
カテゴリ：変光星 | 天文学に関する記事
not-is-a
天体
日本の市町村
恒星
新潟県の市町村
カテゴリ
is-a
is-a
変光星
連星
長岡市
not-is-a
• 爆発変光星
•アメリカ変光星観測者協会
not-is-a
• 長岡まつり
• 北越銀行

is-a関係の判定方法

- 関連手法
  - 文字のパターンマッチでis-a関係を判定
    - 再現率(網羅性)が低い

- is-a関係でないリンクを網羅的に削除
  - 残ったリンクをis-a関係と判定
    - 再現率の向上

Wikipedia のリンクとis-a 関係

- 規則1. 意味が抽象的過ぎる単語の場合はnot-is-a
- 規則2. 親子が意味的に類似していない場合はnot-is-a 関係とする関係とする
- 規則3. 親が固有名詞の場合はnot-is-a 関係とする
- 規則4. 子名の前方が親名と一致する場合はnot-is-a 関係とする

- これらの規則をもとにカテゴリ間のis-a関係を抽出

Wikipedia のリンクとis-a 関係

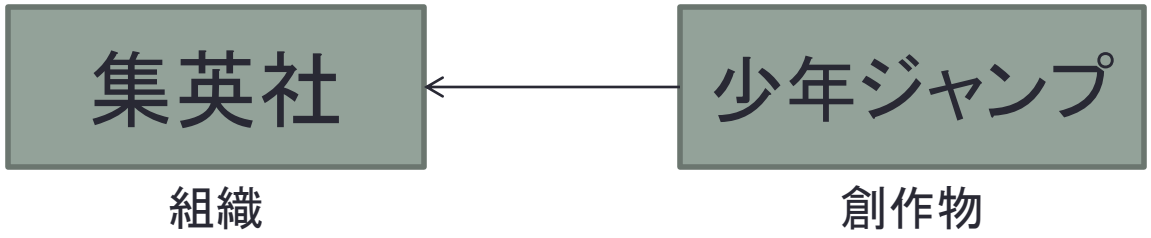
- (例) 技術 ← 道具、社会 ← 経済
- 抽象的な単語は意味が多様
  - 単語間の関係を明確に決定し難い
- 「社会 ← 日本の社会」
  - Is-a関係だが一律にnot-is-a関係としている

Wikipedia のリンクとis-a 関係

- 規則1. 意味が抽象的過ぎる単語の場合はnot-is-a
- 規則2. 親子が意味的に類似していない場合はnot-is-a 関係とする関係とする
- 規則3. 親が固有名詞の場合はnot-is-a 関係とする
- 規則4. 子名の前方が親名と一致する場合はnot-is-a 関係とする

Wikipedia のリンクとis-a 関係

- (例)筆記用具 ← 万年筆メーカー
  集英社 ← 少年ジャンプ



- 9種類の意味属性を設定

人	組織	施設	地名	具体物	創作物	動植物	イベント

- SVM による分類器でカテゴリと記事を分類
  - どの意味属性にも分類されない単語=抽象的
  - 親子の意味属性が違う=意味的に類似していない

Wikipedia のリンクとis-a 関係

- 規則1. 意味が抽象的過ぎる単語の場合はnot-is-a
- 規則2. 親子が意味的に類似していない場合はnot-is-a 関係とする関係とする
- 規則3. 親が固有名詞の場合はnot-is-a 関係とする
- 規則4. 子名の前方が親名と一致する場合はnot-is-a 関係とする

Wikipedia のリンクとis-a 関係

- (例)少年ジャンプ ← ONE PIECE
  新潟県 ← 長岡市
- 固有名詞は基本的に下位に単語持たない
- 多くはpart-of関係である

- 既存の辞書
  - 単語が固有名詞として辞書に登録されているか
- 英語のWikipediaの表記
  - 頭文字が大文字かどうか　(例)The Beatles

Wikipedia のリンクとis-a 関係

- 規則1. 意味が抽象的過ぎる単語の場合はnot-is-a
- 規則2. 親子が意味的に類似していない場合はnot-is-a 関係とする関係とする
- 規則3. 親が固有名詞の場合はnot-is-a 関係とする
- 規則4. 子名の前方が親名と一致する場合はnot-is-a 関係とする

Wikipedia のリンクとis-a 関係

- (例) 火星 ← 火星の衛星、缶 ← 缶コーヒー
  - 子と親はpart-of 関係や話題が類似した関係にあることが多い
- パターンマッチングにより判別可能
- 誤り例　血液 ← 血球

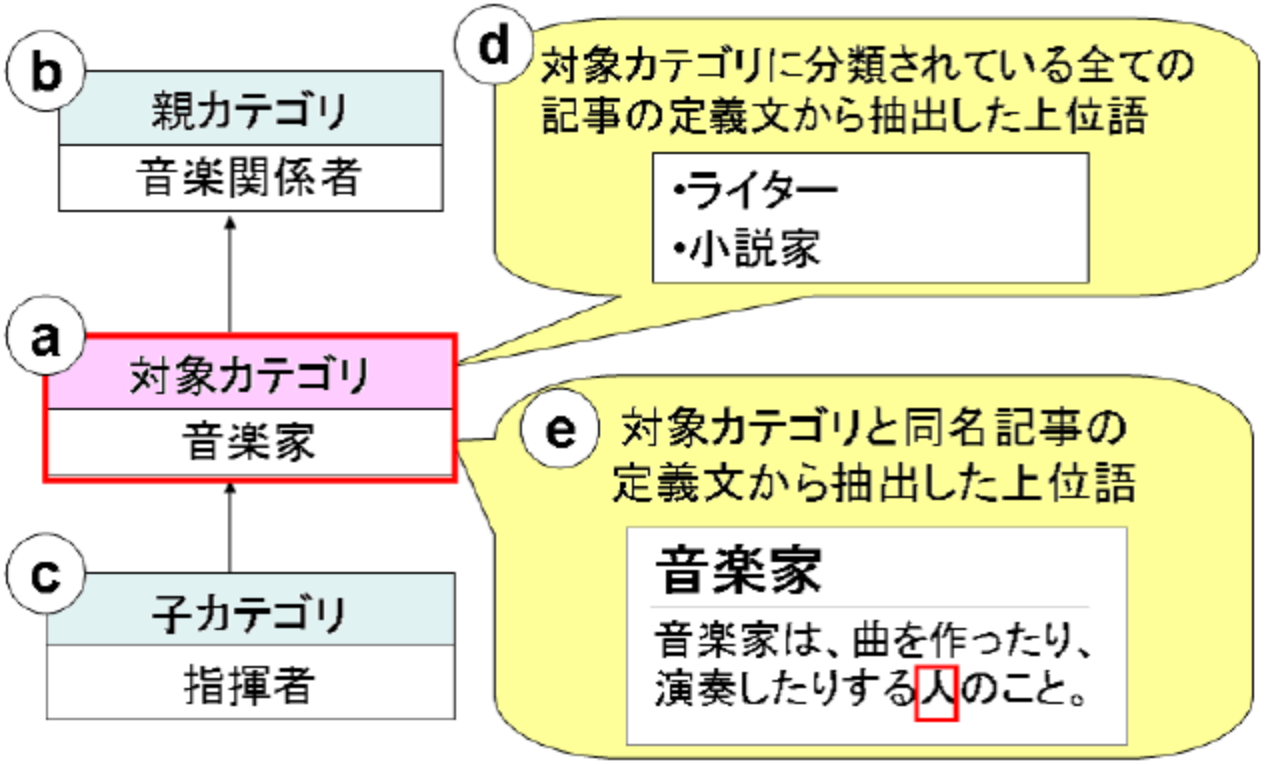
- カテゴリ間で適合率98.9%　再現率99.3%
- カテゴリ-記事間で適合率99.3%　再現率98.9%

カテゴリ分類

- カテゴリ分類のための素性に利用する単語
- a. 対象カテゴリ名
- b. 親カテゴリ名
- c. 子カテゴリ名
- d. カテゴリ中の記事の定義文からとれる上位語
- e. カテゴリと末尾の形態素が一致する記事の定義文からとれる上位語

カテゴリ分類

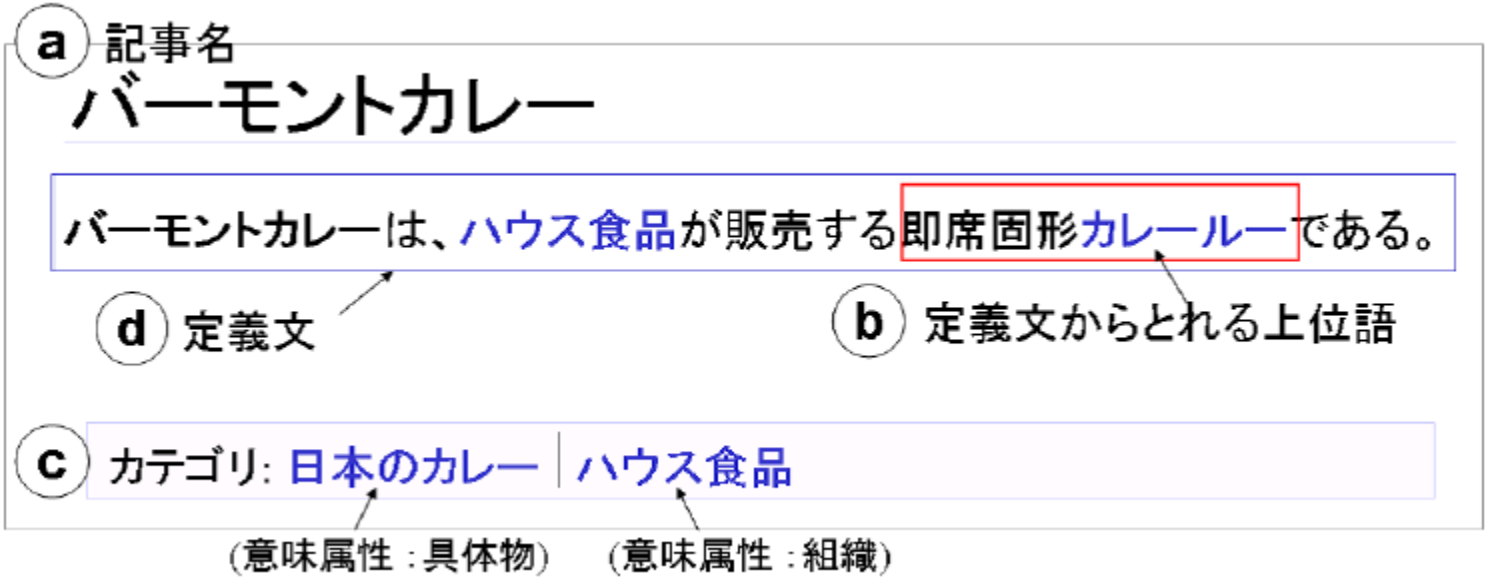
- カテゴリ分類のための素性に利用する単語
- a. 対
- b. 親
- c. 子
- d. カ
- e. カ



からとれる上

記事分類

- a. 対象記事名
- b. 記事の定義文からとれる上位語
- c. 対象記事に付与されているカテゴリ名
- d. 記事の定義文



分類できなかった残り

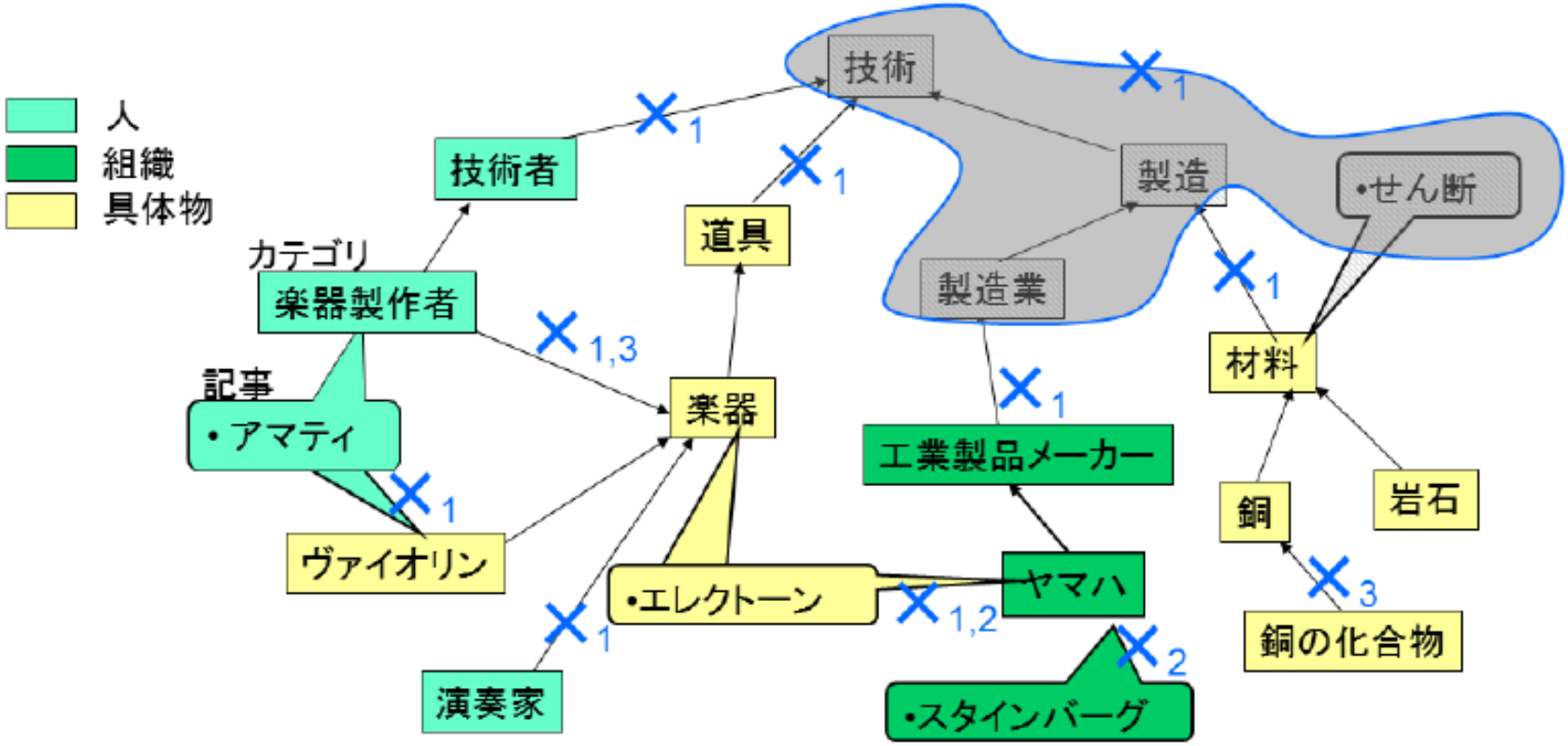
カテゴリ:カクテル （意味属性:具体物）

	記事名	意味属性
既に意味属性が確定した記事	ウーロンハイ	具体物
	サワー	具体物
	ハイボール	具体物
	ホットカクテル	具体物
	バーテンダー	人
意味属性が未確定の記事	カルピスサワー	未確定
	カンパリ・ビア	未確定

“具体物”の割合80%

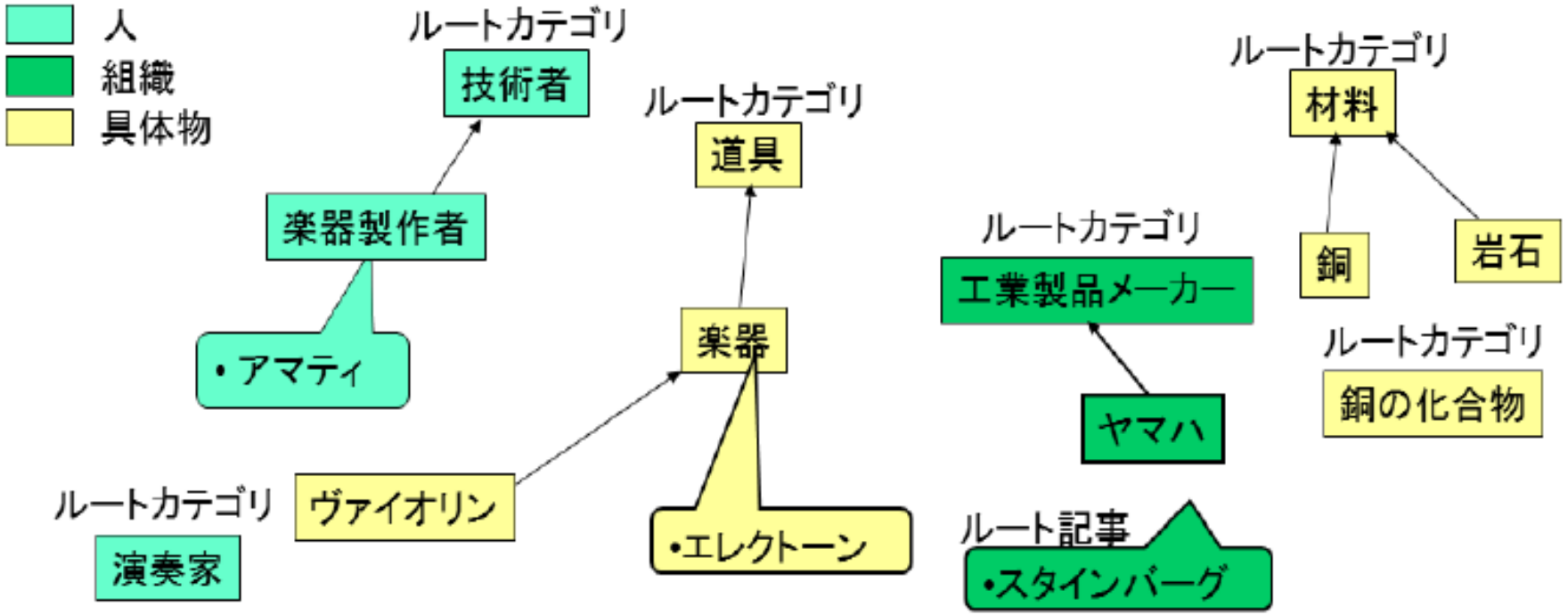
オントロジー階層の再構成

- not-is-a関係と意味属性以外のカテゴリを削除



オントロジー階層の再構成

- 同じ意味属性のカテゴリと記事の階層が構築



オントロジー階層の再構成

- 最上位のカテゴリの下位に同じ意味属性のカテゴリ記事を接続