シソーラスの階層的構造を利用した弱教師あり固有表現抽出

芝原 隆善 [1]　山田 育矢 [1,2]　西田 典起 [1]　寺西 裕紀 [1]　古崎 晃司 [1,3]　松本 裕治 [1]
[1] 理化学研究所　[2] Studio Ousia　[3] 大阪電気通信大学
takayoshi.shibahara@a.riken.jp
{ikuya.yamada, noriki.nishida, hiroki.teranishi}@riken.jp
{kouji.kozaki, yuji.matsumoto}@riken.jp　ikuya@ousia.jp　kozaki@osakac.ac.jp

掲載号の情報

2024 年 31 巻 3 号 pp. 984-1014.
doi: https://doi.org/10.5715/jnlp.31.984

概要

固有表現抽出は，自然言語処理において基本的で重要なタスクである．しかし，大量の教師データを必要とする従来の固有表現抽出は，ユーザーに応じた多様な粒度のカテゴリを抽出するという実社会の需要に柔軟に対応できていない．既知語が出現する文脈を擬似教師データとして利用する弱教師あり固有表現抽出は，大規模なシソーラスと組み合わせることでこの多様なカテゴリの需要に対応できる．弱教師あり固有表現抽出の先行研究は，擬似教師データの誤りに頑健な学習法を提案してきたが，これらの学習法の結果作られたモデルには，関心のあるカテゴリと無関心なカテゴリの境界を超えて予測してしまうという副作用があった．この副作用に対し本研究では，ユーザーの関心のあるカテゴリを含むシソーラスの全カテゴリを擬似教師データ作成に活用する手法を提案し，実験を通じてシソーラスに含まれる総体的な知識の有用性を明らかにした．

参考文献

[1] 芝原隆善, 大内啓樹, 山田育矢, 西田典起, 寺西裕紀, 古崎晃司, 渡辺太郎, 松本裕治. ユーザの興味があるカテゴリに応じた NER システム構築フレームワーク. 言語処理学会年次大会, 浜松, March 2022.

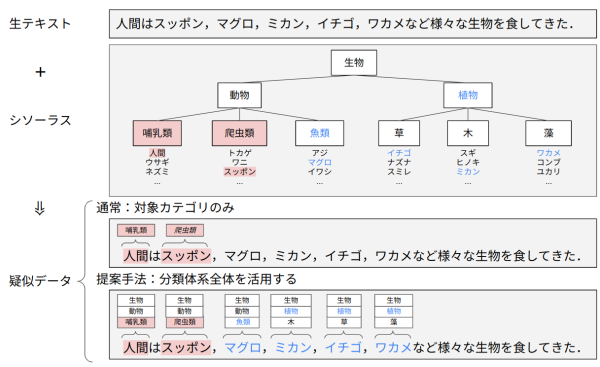

図 1　既存の弱教師あり固有表現抽出との比較。通常の弱教師あり固有表現抽出では関心のあるカテゴリ: 対象カテゴリ（「哺乳類」・「爬虫類」）に含まれる語句のみを辞書マッチによる擬似アノテーションに活用する。先行研究 [1] ではそれらに追加して、対象カテゴリ及びその祖先と兄弟関係にあるカテゴリ: 補完カテゴリ（「魚類」・「植物」）も辞書マッチによる擬似アノテーションに活用する。提案手法ではシソーラスに含まれるシソーラスの全カテゴリを辞書マッチによる擬似アノテーションに活用する。例えば「マグロ」や「ミカン」などのエンティティや「魚類」・「植物」などのカテゴリも擬似教師データに活用する。赤色のカテゴリは対象カテゴリ、青色のカテゴリは補完カテゴリをそれぞれ意味している。