科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２８　日現在

機関番号：　１１３０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１５９

研究課題名（和文）語彙データベースと大規模コーパスに基づく意味クラス推定器の開発

研究課題名（英文）Building Named Entity Recognizers by combining a large-scale lexicon and corpora

研究代表者
岡崎　直観（Okazaki, Naoaki）

東北大学・情報科学研究科・准教授

研究者番号：５０６０１１１８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円、（間接経費）　　1,020,000円

研究成果の概要（和文）：本研究の目的は，テキストから特定の意味クラスに属する概念・実体の表現（例えば製品名や病名など）を抽出するプログラム（意味クラス推定器）を低コストで構築することである．この目的の達成のため，①意味クラスが付与された訓練データの自動獲得，②自動獲得された訓練データからの意味クラス推定器の開発，③意味クラス推定器の性能評価の３つの研究項目に取り組んだ．研究項目①では，辞書に含まれる参考文献情報を利用して学習データの質を向上させる手法を提案し，その効果を実証した．研究項目②③に関して，文脈辞書と呼ばれる意味クラス表現の周辺文脈をマイニングすることで，意味クラスタガーの性能が向上することを確認した．

研究成果の概要（英文）：This research builds Named Entity Recognizers, which extract text mentions of entities or concepts of specific semantic classes (e.g., product names and disease names) from text, at a low cost. In order to achieve this goal, this project addresses three challenges: (1) automatic acquisition of training data with mentions annotated with semantic classes; (2) building Named Entity Recognizers from the automatically acquired training data; and (3) evaluating the Named Entity Recognizers. We proposed a method for improving the quality of automatically acquired training data by using reference information in the dictionary, and demonstrated its effectiveness through the experiments. We also proposed a method for mining context gazetteers, which are dependency paths appearing around expressions of the target semantic classes, and confirmed improvements of accuracy of Named Entity Recognizers.

研究分野：　情報学

科研費の分科・細目：　知能情報学

キーワード：　自然言語処理　固有表現抽出

１．研究開始当初の背景

「魚油の摂取が寒冷暴露に対する耐性を高め，レイノー病の患者の血管けいれんの発現を遅らせる」という例文を考える．医学分野の文献を扱う意味クラス推定器は，この文から「魚油」を物質名，「寒冷暴露」を損傷名，「レイノー病」を病名，「血管けいれん」を症状名，と認識することが望ましい．このような意味解析を大量のテキストに適用すれば，病名と関連性の強い物質名をランク付けしたり，病名とその症状の関係を表す言語パターン（例えば「○○の患者の××の発現」）をマイニングできる．意味クラス推定器は，自然言語処理分野では固有表現抽出器と呼ばれ，テキストから人間にとって有用な知識を掘り起こすための基盤技術である．

意味クラス推定器は，テキストに実体・概念の出現箇所を手作業で付与した訓練データを用意し，サポートベクトルマシンや条件付き確率場などの教師有り学習で構築するのが一般的である．学習により，「○○の摂取」という表現があるとき「○○」は物質名になりやすいとか，「△△の発現」では「△△」は症状名である，というルール群が自動的に獲得・調整され，「魚油」や「血管けいれん」という表現が辞書に無くても，計算機が表現の意味クラスを推定できる．

残念ながら，現状の訓練データの整備は，限られたドメインと意味クラスに限定されている．例えば，CoNLL 2003 や ACE 2005 と呼ばれるデータセットは，新聞記事を対象に人名・会社名・地名などを付与しており，NLPBA や BioCreative II Gene Mention というデータセットは，生命・医学系の文献を対象に，遺伝子名・タンパク質名を付与している．したがって，ドメインが異なるテキスト（Web や会話文など）や，異なる意味クラス（病名，物質名，化合物名，製品名，店名など）に対応した意味クラス推定器を構成するには，人間の手間と労力をかけ，訓練データを準備しなければならない．今後，知識獲得の応用範囲を様々なドメイン・意味クラスに拡張する際，訓練データの入手性が，意味クラス推定器のボトルネックとなる．

一方で，実体・概念の表現事例を収録している語彙データベースは，入手が比較的容易である．代表的なものとしては，UMLS Metathesaurus（生命･医学分野），Wikipedia（カテゴリを意味クラスと見なすことができる），Freebase（一般ドメイン）などが挙げられる．

２．研究の目的

本研究は実体・概念の表現事例から，高性能な意味クラス推定器を開発する．具体的には，以下の研究項目に取り組む．

(1) **意味クラスが付与された訓練データの自動獲得**: Web 文書や論文抄録など，意味クラスが付与されていない大量の生テキストデータから，既存の語彙データベースに含まれる表現を見つけ出し，意味クラスを付与する．テキストマッチングだけの簡単な処理に見えるが，表記揺れと曖昧性という重要な課題を解決する必要がある．

(2) **自動獲得された訓練データからの意味クラス推定器の構築**: 自動獲得した訓練データに対し，半教師有り学習手法を適用し，自動獲得した訓練データの拡充，及び意味クラス推定器の構築を同時に行う．研究項目(1)で獲得した訓練データは，語彙データベースに基づいて意味クラスを付与しているため，語彙データベースの網羅性がそのまま訓練データの網羅性に反映される．語彙データベースの網羅性は完璧ではない（完璧であれば(1)の辞書引きだけで意味クラス推定器が出来てしまう）ため，(1)で構築された訓練データでは，意味クラスに属する実体・概念の表現に対し，意味クラスが正しく付与されない問題が発生する．この問題に対処するため，(1)で自動構築した訓練データに対し，半教師有り学習を適用し，自動獲得した訓練データのエラーを訂正しながら，意味クラス推定器の学習を行う．本研究項目では，訓練データのエラーを訂正するのに有効なモデル･素性の設計，大規模な訓練データから意味クラス推定器を効率良く学習する方法を探求する

(3) **意味クラス推定器の性能評価**: (2)で構築した意味クラス推定器が，与えられた表現事例の意味クラスの表現を過不足なく認識できるか，評価実験を行う．

３．研究の方法

(1) 意味クラスが付与された学習データの自動獲得

概念･実体の表現事例（語彙データベース）に基づき，学習データを自動獲得する手法の設計と実装を行う．研究の目的で述べたように，語彙データベースを参照しながら生テキストの表現に意味クラスを付与する際には，表記揺れと曖昧性の問題を解決する必要がある．研究代表者は表記揺れに関して高速かつスケーラブルな類似文字列検索アルゴリズムを開発した．また，曖昧性に関しても，申請者はテキスト中で用いられている用語の意味を推定する手法を開発した．これらの研究成果を活かし，意味クラスの自動付与における表記揺れと曖昧性の問題を解決する．

このアプローチの有効性を検証するため，CoNLL 2003 や BioCreative III Gene Mention など，既存の訓練データのテキスト部分に対して，人手で付与された正解と比較しながら，本研究項目の手法の設計・実装を行う．

(2) 自動獲得した訓練データから高精度な意味クラス推定器を構築

研究項目(1)で自動獲得した訓練データを出発点とし，半教師有り学習を用いて，意味クラス推定器を構築する．本研究項目のポイントは，語彙データベースが網羅していない表現（false negative）と，自動構築した訓練データが元々の意味クラスから乖離する問題（意味ドリフト;false positive）への対処である．前者については，語彙データベースが収録している表現の周辺の文脈を学習し，例えば「○○の摂取」という文脈があるとき，「○○」は物質名になりやすいというルールを学習し，このルールに当てはまる表現を生テキスト中から自動的に拡充する．一方,「○○の影響」という文脈に対して,「○○」は物質名，病名，症状，治療行為など，様々な意味クラスの表現が用いられると考えられる．そこで，表現の自動拡充を行う際に，その表現の意味クラスに関する曖昧性を推定し，後者の意味ドリフト問題に対処する．

(3) 意味クラス推定器の性能の評価

生テキストに意味クラスを付与する作業を行い，新しいドメインのテキスト・意味クラスに対して，提案手法と従来手法のアプローチを比較する実験を行う．意味クラスとしては，語彙データベースの入手性などを考慮しつつ，従来の訓練データでは採用されていなかった意味クラスを選ぶ．テキストのドメインとしては，多種多様な実体・概念が記述されている生命・医学分野の論文抄録と，本研究の今後の応用が見込まれる Web 文書を予定している．

４．研究成果

Unified Medical Language System (UMLS) の遺伝子名を概念・実体の表現事例（語彙データベース）と見なし，PubMed の論文抄録を生テキストコーパスとして，学習データの自動獲得を行った．具体的には，PubMed の論文抄録のテキスト中に含まれるトークン列が，UMLS に遺伝子名として収録されている場合，該当部分を遺伝子名の正例とした．Gene or Gene Products (GGP) の意味クラスを人手で付与した評価データを用い，UMLS の辞書マッチングによる意味クラス推定の性能を測定したところ，精度 92.1%，適合率 39.0%，再現率 42.7%，F1 スコア 40.8 が得られた．適合率・再現率ともに低く，概念・実体の表現事例と生テキストコーパスの辞書マッチングを行うだけでは，ノイズ（偽正例と偽負例の両方）が多いことが分かった．

このように自動獲得した訓練データを用いて，条件付き確率場（CRF）で意味クラス推定器を構築した．この際，各意味クラスに属するトークンを丸覚えしてしまうことを

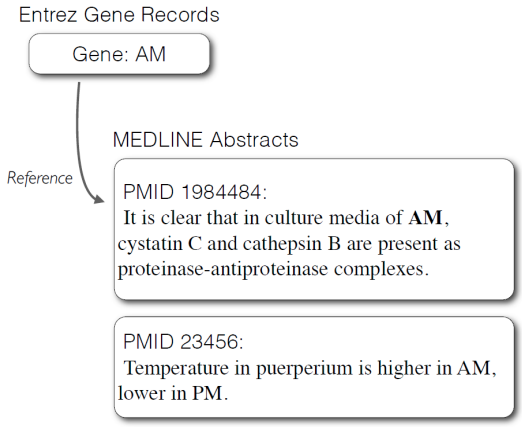

図１ 参考文献情報を利用した学習データ作成

防ぐため，トークンの周辺の単語の出現状況（Ｎグラム）や，トークンの文字種（大文字で始まるとか，すべてが大文字になっている等）を素性として採用した．PubMed 全体に対して教師データを自動獲得し，意味クラスタガーの性能を GGP コーパスで測定したところ，精度 85.8%，適合率 10.2%，再現率 23.8%，F1 スコア 14.3 であった．意味クラスタガーの性能を改善させるため，正例を抽出する際，UMLS レコードの参考文献情報と抄録の文献 ID のマッチングを行うように工夫（図１）したところ，精度 93.7%，適合率 69.3%，再現率 39.1%，F1 スコア 50.0 まで改善が見られた．この性能は，単なる辞書引きによる性能を上回っており，語彙辞書と生テキストから教師有り学習を用いて意味クラスタガーを構築することの意義が示された（学会発表文献７）．

その他の評価対象ドメインとして，日本栄養士会が東日本大震災時に支援活動を行った際の報告書（自由記述）を採用し，そのテキストに意味クラスを手作業で付与した．意味クラスとしては，1.他団体との連携，2.管理栄養士・栄養士の本務としてのサービス（栄養相談など）提供，3.栄養指導・支援（炊き出しの手伝いなどの調理に関するもの），4.事務処理（カルテや支援物資の整理など），5.支援物資・提供者，6.活動場所（避難所，自宅，仮設住宅など）を採用した．このドメインのテキストに対して，意味クラス推定器を構築したところ，約７割の精度であった．低コストで意味クラス推定器が構築できたことから，新しいドメインのテキストや新しい実体・概念に対して，本プロジェクトで研究を進めた手法が有効的であることが実証できた（雑誌論文５）．

さらに，意味クラス推定器の性能向上に関する新しいアイディアの提案も行った．意味クラス推定器の性能が向上させるには，認識したい意味クラスの表現事例を大量に集めた辞書を構築し，現在解析している表現がその辞書に含まれるかどうかを考慮することが効果的であると知られている．この特徴は，

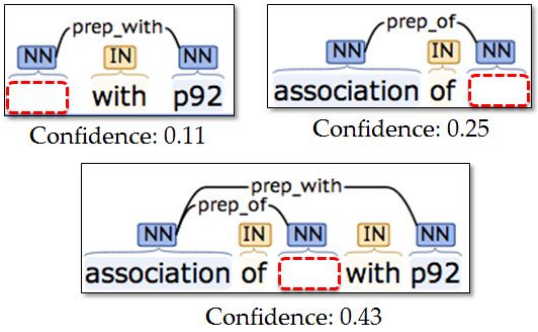

図２ 得られた文脈辞書の例

本研究で構築されている意味クラス推定器にも採用されているが，本研究で構築した意味クラス推定器のエラーを解析していたところ，単語の文脈の情報を増強することで提案手法の性能を改善できる可能性が判った．

具体的には，「文脈辞書」と呼ばれる意味クラスの表現事例とよく共起する文脈のリストを大域的文脈として用いることを提案した．文脈辞書は本研究の手法と同様に，タグ付けされていないテキストコーパスと表現事例辞書を組み合わせ，意味クラスの表現と係り受け関係にある表現（係り受けパス）をマイニングすることによって構築した．生命医学分野を対象とした実験では，述語項構造関係のような高い確信度を持つ文脈辞書を構築出来たこと（図２），本研究で構築している意味クラスタガーの性能をさらに向上できることを確認した（学会発表文献２）．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　６件）

1. Han-Cheol Cho, Naoaki Okazaki, Makoto Miwa, Jun'ichi Tsujii. Named entity recognition with multiple segment representations. Information Processing & Management, Vol. 49, No. 4, pp. 954–965, 2013.（査読あり）
DOI: 10.1016/j.ipm.2013.03.002
2. Xu Sun, Naoaki Okazaki, Junichi Tsujii, Houfeng Wang. Learning Abbreviations from Chinese and English Terms by Modeling Non-local Information. ACM Transactions on Asian Language Information Processing, Vol. 12, No. 2, pp. 5:1–5:17, 2013.（査読あり）
DOI: 10.1145/2461316.2461317
3. 鍋島啓太，渡邉研斗，水野淳太，岡崎直観，乾健太郎．訂正パターンに基づく誤情報の収集と拡散状況の分析．自然言語処理，Vol. 20，No. 3，pp. 461–484，2013 年．（査読あり）
DOI: 10.5715/jnlp.20.461
4. 高瀬翔，岡崎直観，乾健太郎．カテゴリ間の兄弟関係を活用した集合拡張．自然言語処理，Vol. 20，No. 2，pp. 273–296，2013 年．（査読あり）
DOI: 10.5715/jnlp.20.273
5. 岡崎直観，鍋島啓太，乾健太郎．言語処理による分析 – 日本栄養士会活動報告の分析．日本栄養士会雑誌，Vol. 55，No. 12，pp. 6–8，2012 年．（査読なし）
6. 岡崎直観，辻井潤一．集合間類似度に対する簡潔かつ高速な類似文字列検索アルゴリズム．自然言語処理，Vol. 18，No. 2，pp. 89–118，2011 年．（査読あり）
DOI: 10.5715/jnlp.18.89

〔学会発表〕（計　７件）

1. 佐藤貴大，岡崎直観，乾健太郎．ウェブ文書の構造を利用した場所名・住所ペアの獲得．第 27 回人工知能学会全国大会（JSAI2013），富山国際会議場（富山県），2013 年 6 月 4 日～7 日．
2. Han-Cheol Cho, Naoaki Okazaki, Kentaro Inui. Inducing Context Gazetteers from Encyclopedic Database for Named Entity Recognition. Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013), pp. 378–389, Gold Coast, Australia, April 14-17 2013.
3. Han-Cheol Cho, Naoaki Okazaki, Kentaro Inui. Exploiting Dependency Context Gazetteers for Named Entity Recognition. 言語処理学会第 19 回年次大会(NLP2013)，pp. 220–223，名古屋大学（愛知県），2013 年 3 月 13 日～15 日．
4. Sho Takase, Naoaki Okazaki, Kentaro Inui. Set Expansion using Sibling Relations between Semantic Categories. Proceedings of the 26th Pacific Asia Conference on Language,Information and Computation (PACLIC 26), pp. 567–576, Bali, Indonesia, November 9 2012.
5. 高瀬翔，岡崎直観，乾健太郎．名詞カテゴリからの関係知識獲得に向けて．NLP 若手の会 第 7 回シンポジウム，東北大学（宮城県），2012 年 9 月 3 日～4 日．
6. 高瀬翔，岡崎直観，乾健太郎．意味カテゴリの階層関係を活用した集合拡張．言語処理学会第 18 回年次大会(NLP2012)，pp. 475–478，広島市立大学（広島県），2012 年 3 月 14 日～16 日．
7. Yu Usami, Han-Cheol Cho, Naoaki Okazaki, Jun'ichi Tsujii. Automatic Acquisition of Huge Training Data for Bio-Medical Named Entity Recognition. Proceedings of BioNLP 2011 Workshop,

pp. 65–73, Portland, Oregon, USA, June 23 2011.

〔図書〕（計　0件）

〔産業財産権〕
○出願状況（計　0件）

○取得状況（計　0件）

〔その他〕
無し

６．研究組織
(1)研究代表者
岡崎　直観（OKAZAKI, NAOAKI）
東北大学・大学院情報科学研究科・准教授
研究者番号： 50601118

(2)研究分担者
無し

(3)連携研究者
無し