様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　２９　日現在

機関番号：　１２６０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２８７３

研究課題名（和文）形式の異なる語彙知識の相互運用の試み

研究課題名（英文）Towards integrated use of different lexical knowledges

研究代表者
加藤　恒昭（Kato, Tsuneaki）

東京大学・大学院総合文化研究科・教授

研究者番号：６０３３４２９９

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：単語の意味に関する知識を表現する手法はいくつかあるが，本研究ではそれらの相互関係を明らかにし，異なる手法を総合的に用いる手法を検討した．意味の包含関係（包摂関係）を中心した表現については，既存の語彙知識において，その関係の認定に恣意性が高く，複数の知識を融合するのは難しいという否定的な結論を得た．一方，巨大なテキストから自動獲得される語彙知識について，これまでの語の知識から，語の対の知識へと拡張することで，意味の類似性だけなく包摂関係を含め様々な意味関係を表現できるようになることを示した．

研究成果の概要（英文）：These are three major frameworks to represent lexical knowledge, knowledge on word senses. In this study, we examined relationships among those and investigated their integrated use. On the representation based on semantic subsumption (hpyonymy) relations, we examined existing lexical knowledges, and reached a negative conclusion that arbitrary decision criteria of subsumption relations make their integrated uses difficult.
On the other hand, on the lexical knowledge that can be constructed automatically from large text corpus, we extended those from word representation to word pair representations and allowed those to represent not only semantic similarities of words but also several semantic relations including subsumption relations.

研究分野：　自然言語処理

キーワード：　語彙意味論　語彙オントロジ　概念体系　包摂関係　分散表現　意味関係認識

１．研究開始当初の背景

語の意味（語義）に関する知識である語彙知識は，その表現や獲得が言語学的な興味関心であるだけでなく，自然言語処理や情報アクセス等での利用が期待され，工学的応用の重要性も高い．人間向けの辞書での自然言語による記述を除けば，語彙知識は構造表現（語彙分解に基づく語彙知識），ネットワーク表現（語彙オントロジ），分散表現（素性空間表現）という３種類に分類される．構造表現は語をより細かい意味要素からなる構造として表現し精密な意味の表現が期待できるが，これを用いた大規模な語彙知識は構築されていない．ネットワーク表現は包摂関係（上位下位関係）など，語の意味関係を表現したもので，日本語においても EDR 辞書，日本語 WordNet, 述語項構造シソーラスなど大規模な語彙知識が構築されている．分散表現は巨大コーパスから自動的に獲得できるが，意味的類似性以上のものを表現できるかは明らかでない．このようにこれらは，それぞれに利点と課題を有しており，それらの相互関係やそれらを融合させる仕組みは明らかでなかった．

２．研究の目的

異なる形式で表現された語彙知識を融合し知識の相互運用を可能とすることを目的とし，それらの相互関係を明らかにする．既存語彙知識の調査を通じて，異なる形式での表現に含まれている知識がどのように関連するかという観点から，その特徴や問題点を明らかにするとともに，それらの問題を解決する語彙知識の表現を提案する．

３．研究の方法

大きく二つの方法で研究を進める．
(1) 実用的な規模での既存語彙知識であるネットワーク表現（語彙オントロジ）を調査分析する．ネットワーク表現の背骨をなす意味関係である包摂関係について，ネットワーク表現による複数の語彙知識を対応づけて比較することで，体系の客観性や適切性を検討する．あわせて構造表現との相互関係の観点から，動詞語義の構造表現において重要な素性である自他の対立を手掛かりとして包摂関係の体系について検討する．ここでは．構造表現において豊かな情報を持つ動詞を中心とした用言に着目している．
(2) 分散表現が，包摂関係などネットワーク表現に記されている意味関係を表現しているかを明らかにするために，既存の分散表現から意味関係を導く手法を検討する．さらにそのような意味関係をより適切に表現するように分散表現を拡張することを検討する．

４．研究成果

(1) 既存ネットワーク表現知識の中でもっともよく用いられる EDR 辞書に着目し，相互運用を意識して，その特徴分析を行った．EDR 辞書は，日本語単語と英語単語の語義が総合的に表現されているという多言語性を持つが，実際に調べてみると，日本語単語に関連する概念（語義と直接語義とはなっていない上位の意味を合わせて概念と呼ぶ）と英語単語が関連する概念との重なりは小さく，日本語動詞の語義と関連する概念は全体の 40% 強しかない（日本語動詞と英語動詞が完全に分離していて重なりがないとしても 50%となるはずで，この 40%という値は非常に小さい），さらに日本語動詞と英語動詞の共通の語義となっている概念は日本語動詞語義全体の 14%しかない等，両言語の概念体系（包摂関係のネットワークで，木の拡張である DAG の形を持つ）の融合に必ずしも成功していないことを明らかにした．また，動詞語義の上位概念が［事象］ではなく，［物事］に繋がっているという特殊な構造が存在することも発見し，その原因が，動詞や名詞等の分類に関わりなく概念階層が構築されているという品詞を越えた語義記述であることを示した．このような概念の構造は無駄であるとともに誤った推論を導くことから問題であり，その一部を抜き出して活用するのが妥当であることを明らかにした．

(2) 前述の観点でその一部を抜き出した EDR 辞書について，日本語 WordNet との相互運用のために，両者の概念のアライメント（対応づけ）を試み，有意味な対応関係が得られるかを調査した．抜き出した EDR 辞書は動詞語義が［事象］を最上位概念とする一つの DAG にまとめられている．この DAG の 4 段目の 221 概念を対応づけの対象とした．一方，日本語 WordNet の動詞概念は比較的浅い DAG の森を構成している．この森は 15 に大分類されており，全体で 559 の DAG からなる．それぞれの DAG の根である 559 概念を対応づけの対象とした．対応づけはその概念を意味とする語の重複に基づいて行った．つまり，同じ動詞の語義となっている概念どうしは対応づけられるものとした．ただ，このような対応はひとつの語が複数の語義を持つために自明ではない．ここに統計的機械翻訳で用いられる単語アライメントの手法を用いた．対応づけ結果を考察して明らかになったことは，このレベルの概念どうしの対応づけ自体も充分には明確ではないが，それでも適切と思われる場合も少なくないこと，むしろ問題であるのはそれら対応づけられた概念の上位分類（大分類）が大きく異なることであった．これは，一方のある概念が他方の概念体系のまとまった部分として対応づけられることが少ないということで，それらの設計指針が異なることを意味する．このことは概念体系

構築の恣意性を示唆し，複数の語彙知識の統合が困難であることを示している．
(3) 動詞語義にはいくつかの分類の観点がありうるので，いわゆる「正しい」概念体系の構築はないとしても，それを特徴付けたり，それが一貫しているかを検証したりするような分析は可能であると考え，そのような分析に別の意味関係との関係を用いることを提案した．つまり，同じ意味関係にある語義の対は，同じ包摂関係のトポロジを持つべきであるという制約であり，ある意味関係がどのような包摂関係のトポロジと対応づけられるかでその語彙知識を特徴付けられると考えられるし，一貫した対応づけがないとすれば，その設計も一貫したものではないと言える．動詞語義の場合，そのために用いる意味関係はいくつか考えられるが，本研究では，自動詞・他動詞の対立を取り上げた．構造表現による語彙知識では，他動詞語義は対応する自動詞語義を下位構造とし，それにそれを引き起こす行為に相当する上位構造を持つとされる．例えば，何かを「開ける」ことは，対象に働きかけてその対象が「開く」ようにすることであるとされる．このような関係が概念体系にどう反映されているかを調査した．
　既存ネットワーク表現知識の EDR 概念辞書と述語項構造シソーラスを比較することを考え，それらの語義記述注釈づけを行い，両者に共通（一対一対応）し，かつ自他対立の関係にある語義の対(165 対)を選び出した．それらのネットワーク構造中での位置関係を分析した．EDR 概念辞書は他動詞語義の配置を上位構造に基づいて行っている（自動詞語義と他動詞語義は根の近くで大分類される）のに対し，述語項構造シソーラスは下位構造に基づいて行っている（自動詞語義と他動詞語義は同じもしくは非常に近い分類に位置する）という設計指針の違いを明らかにしたことに加え，EDR 概念辞書のひとつの特徴である多重継承（複数の上位概念を持つこと）が誤った推論を導くことを示した．また，知識構築における作業ミスの発見など，異なる形式での表現を付き合わせることの有効性を示した．一方で，いわゆる自他対立が必ずしも一つの意味関係と対応していないなど方法の問題点も明らかとなった．

(4) 分散表現と意味関係の関係については，分散表現が包摂関係などネットワーク表現に記されている意味関係を表現しているかを明らかにするために，既存の分散表現から意味関係を導く手法を検討することが課題であったが，これについては，関連研究により既存の分散表現だけからでは不十分であるとの結論が出た．そのため，本研究では，既存の分散表現から導けるものは何か，意味関係をより適切に表現するような分散表現の拡張はいかなるものかを明らかにすることを課題とした．

　語の意味の分散表現は，その語がどのような語の近傍で用いられるかという情報を基に，巨大なコーパスを利用して作成される．近年では，ニューラルネットワークを用いて，近傍に出現する語を予測するようにパラメータが調整（学習）され，そのパラメータを用いて語からその語の分散表現への写像が行われる．表現は数百次元の数値ベクトルであり，このベクトルどうしのコサイン尺度が，同じような文脈に現れやすいという意味での語の意味の類似性に対応することが知られている．
　分散表現が包摂関係などネットワーク表現に記されている意味関係の情報を有しているかという問題は，二つの語の分散表現が与えられた際にその語が包摂関係にあるかを判定する課題として定義される．語の分散表現（ベクトル）の差や連結をデータとして，包摂関係を学習することができるかが検討されたが，十分には不可能で，高々それぞれの語の上位語下位語としての典型性が学ばれているにすぎない，さらにそれも適切には学ばれていないとの報告がなされた．
本研究はこの点について，語の上位語下位語としての典型性をより適切に学習する方法を提案した．学習のためのデータが包摂関係が木に近い DAG であることから，下位語に比べ同じ上位語の学習データでの出現回数が多いことが正しい学習を妨げている原因であると洞察し，学習データを間引くことで，分類の性能が向上することを複数のテストデータで示した．

(5) 分散表現の拡張として，包摂関係などを導くために，従来の語の分散表現に加えて，語の対の共起パタンを分散表現化した知識が有益であることを示した．従来の語の分散表現がその語の近傍に出現する語を予測するように学習されるのに対し，語の対の分散表現は二つの語が文内に共起する際のその間のパタン（単語列や依存構造）を予測するよう学習される．語の対の表現に着目した既存研究はあるが，ここではニューラルネットワークを用いることで，データが疎らであることに起因する問題を回避し，一般性の高い表現を得ることができることが大きな利点である．このような語の対の分散表現を用いることで，包摂関係を含めた語の意味関係の認識性能は大きく向上し，ネットワーク表現で記述されている意味関係を含んだ分散表現への足掛かりを得た．

５．主な発表論文等

〔雑誌論文〕（計　1　件）

① 林 良彦．言語学と AI，人工知能学会誌，査読無（依頼有）Vol.32，No.3，pp.384-393，2017．

〔学会発表〕（計　12　件）

① Koki Washio and Tsuneaki Kato. Filling Missing Paths: Modeling Co-occurrences of Word Pairs and Dependency Paths for Recognizing Lexical Semantic Relations. NAACL HLT 2018, 2018 年 6 月 10 日, New Orleans, USA.

② Koki Washio and Tsuneaki Kato. Under-sampling Improves Hypernymy Prototypicality Learning. LREC2018, 2018 年 5 月 11 日, 宮崎.

③ 李 凌寒, 加藤 恒昭. FrameNet を利用した談話関係の認識. 第 24 回言語処理学会年次大会(NLP2018), A2-1, 2018 年 3 月 13 日, 岡山.

④ 鷲尾 光樹, 加藤 恒昭. 単語ペアと依存構造パスの共起モデリングを用いた語の意味関係の分類. 第 24 回言語処理学会年次大会(NLP2018), B4-4, 2018 年 3 月 14 日, 岡山.

⑤ 加藤 恒昭. 動詞語義の階層的分類に関する一考察. 第 24 回言語処理学会年次大会(NLP2018), B4-5, , 2018 年 3 月 14 日, 岡山.

⑥ 金田 健太郎, 小林 哲則, 林 良彦. 語義・概念の分散表現を利用した Semantic Taxonomy Enrichment. 第 24 回言語処理学会年次大会(NLP2018), P7-10, 2018 年 3 月 14 日, 岡山.

⑦金田 健太郎, 小林 哲則, 林 良彦. 語義・概念の分散表現を利用した単語間の意味関係分類, 第 23 回言語処理学会年次大会(NLP2017), P5-2, pp.214-217, 2017 年 3 月 14 日, つくば.

⑧ Kentaro Kanada, Tetsunori Kobayashi, Yoshihiko Hayashi. Classifying Lexical-semantic Relationships by Exploiting Sense/Concept Representations, EACL2017 Workshop on Sense, Concept and Entity Representations and their Applications, pp.37-46, 2017年4月4日, Valencia, Spain.

⑨ 鷲尾 光輝, 加藤 恒昭. 分散表現を用いた語の上位下位関係の学習 ― Lexical Memorizationの緩和― 第23回言語処理学会年次大会(NLP2017) B6-4 pp. 887-890. 2017年3月16日, つくば.

⑩ 加藤 恒昭, 林 良彦. 日本語動詞に関する EDR 概念辞書の分析, 研究報告自然言語処理（NL） 2015-NL-224(12),1-10, 2015 年 12 月 4 日, 名古屋.

〔図書〕（計　1　件）

① 加藤 恒昭, 菊井 玄一郎, 林 良彦, 森 辰則（共訳） 統計的自然言語処理の基礎, 共立出版 pp. 606, 2017.

６．研究組織

(1)研究代表者
加藤 恒昭（Kato Tsuneaki）
東京大学・大学院総合文化研究科・教授
研究者番号：60334299

(2)研究分担者
林 良彦（Hayashi Yoshihiko）
早稲田大学・理工学術院・教授（任期付）
研究者番号：80379156