審 査 の 結 果 の 要 旨

氏名　ボッレーガラ　ダヌシカ　タルパティ

本論文は「A Study on Attributional and Relational Similarity between Word Pairs on the Web（ウェブ上での単語対間の属性類似性と関係類似性に関する研究）」と題し，英文８章から成る．

第１章「Introduction（序論）」では，主テーマである属性類似性(attributional similarity)と関係類似性(relational similarity)について，それらの定義と差異，類似度の計測法などの関連事項を記している．類似性は人工知能，自然言語処理，認知科学，哲学といった様々な分野に広がる概念であり，属性類似性と関係類似性として大きく２つに分類することができる．属性類似性は２つの概念が持つ属性集合間の類似性が高ければ高くなる．一方，関係類似性は２つの単語対の間に成り立つ意味的関係の近さを評価する．例えば，単語対（ダチョウ，鳥）と単語対（ライオン，猫）を考えると，ダチョウは地球上で存在する最大の鳥であり，またライオンは最大の猫（科の動物）である．従って，それぞれの単語対に含まれる２単語間では「～は最大の～である」という関係が成立し，この例の単語対の間には高い関係類似性が存在することになる．

第２章「Semantic Similarity（意味的類似性）」では，Web 検索エンジンを利用する属性類似性の計測手法を提示している．最初に既存手法をまとめている．辞書における単語の階層的分類を利用する手法とは別に，検索エンジンを用いるものとして２単語の AND 検索によるヒット件数に基づく手法，検索エンジン出力のスニペットから得られる当該２単語を含む単語列によって判別する手法を紹介している．これに対し，本論文では前者をグローバルな特徴の利用，後者をローカルな特徴の利用と位置付け，両者の特徴を合わせることで考案した新手法を記している．ここでローカルな特徴としてスニペットに表れる当該２単語を含む大量の単語パターンを利用するのだが，類似性判定に有効な単語パターンを機械学習により求める有効な手法を導入している．提案手法の有効性は，WordNet から得られる 5000 の類義単語対データ，人手で判定された Miller-Charles の 30 単語対ベンチマーク・データを用いた実験により，既存手法を上回る判別性能が得られることを示している．また，本提案手法は人間のコミュニティ抽出や単語の曖昧性解消のタスクにおいても有効に利用できることを示している．

第３章「Relational Similarity（関係類似性）」では，単語対間の関係類似性を Web 検索エンジンを利用して計測する新手法を提示している．関係類似性を計測するためには，まず与えられたそれぞれの単語対の２単語間にどのような関係が存在するかを知る必要があるが，そのために Web 検索エンジンが返すスニペットを用い，関係を表す２単語を含む周辺文脈単語パターンを抽出する．まず関係類似性の判定に有効な単語パターンを，類似関係の単語対（正例）と類似関係性のない単語対（負例）の集合を用いて分割表のカイ２乗検定により絞り込む．一つの関係を表現する単語パターンが複数存在し得るため，単語パターンの分布類似度を用いて単語対の２単語間の関係を特徴付ける必要があるが，これを効果的に行うために意味的に近い単語パターンを単語対に対する分布形の近さに基づきクラスタリングする．この段階でも 10 万を超える膨大な数の単語パターンを扱うことになるため，効率的なクラスタリング法を考案し，導入している．このクラスタリングはデータのスパースネス問題を軽減する効用ももたらす．最終的に生成された単語パターンのクラスタを特徴量として，単語対を特徴ベクトルで表し，２つの単語対間での類似性を計測するのだが，この特徴量間には相関が存在し独立な特徴量とはならないので，この特徴量間の相関を考慮するマハラノビス距離尺度を用いて計測する．このマハラノビス距離尺度計算に必要なマハラノビス行列は，関係類似性に関する正例と負例の集合から学習できることも示している．

以上の提案手法の性能は，SAT 問題（Scholastic Assessment Test：与えられた単語対の単語間の関係と近い関係を持つ単語対を４候補の中から選択する問題で，英語が母国語の高校生の平均正答率は 57%）と，５種の関係（企業の買収関係，人－生誕地関係，企業－CEO 関係，企業－本社所在地関係，人－専門分野関係），のいずれかを持つ 100 単語対データを対象とする実験により実証しており，既存手法よりも優れた性能が得られることを示している．後者の実験の場合，単語パターンクラスタ数は約１万となっている．既存手法の中で優れているのは P. D. Turney による手法

であるが，これは後者の実験課題の場合に数日オーダの処理時間を要する．これに対し提案手法は数時間になるといったように，計算時間の点でも優れていることを示している．

第４～５章はエンティティ間の類似性に関係する Web における同姓同名人物判別問題，別名検出問題，人物の属性抽出問題を扱い，それぞれ新手法を提示している．

第４章「Personal Name Disambiguation（人名曖昧性解消）」では，Web における同姓同名人物（英語で namesake と称される）判別を行う新手法を提示している．この手法では特定の人物に対して関係が深い複数キーワードを列挙するが，このキーワード候補は人物名で検索された Web ページのテキストより，C-value 法と固有表現認識(named entity recognition)技術により求める．そして，人名を検索語として Web 検索エンジンから検索結果として得られる各 Web ページのスニペットに現れる上記キーワード候補の出現分布形の情報を基にして，ボトムアップ凝集型クラスタリングを行うことにより，同一人物の Web ページを同一クラスタに纏める．同姓同名人物判別法は色々提案されているが，実験により提案手法の優位性を示している．

第５章「Name Alias Detection（別名検出）」では，同姓同名とは逆に同一エンティティを指す複数の別名を Web から検出する新手法を提示している．例えば，“松井秀樹”は“ゴジラ”とも呼ばれることがあり，このゴジラが松井秀樹の別名となる．ここで，人物名の別名だけでなく，地名等の別名も対象になる．考案した検出法で用いる主な特徴は，検索エンジン出力スニペットに現れる正式名－別名の多数の周辺文脈単語パターン，正式名 AND 別名候補の検索ヒット数（グローバルな特徴量），更に同一 URL を指すアンカテキスト（これらは同一エンティティの呼称であることが多い）である．上記の単語パターンは正しい正式名－別名の対を与えて，有用なもの 200 パターンを抽出して特徴量として用いている．これらの特徴量を用いる SVM を学習する．そして別名が未知の正式名を与えると，スニペットから別名候補を抽出し，上記 SVM により別名の識別を行う．提案手法の性能は実験により既存手法より優れていることを示している．

第６章は「Attribute Extraction（属性抽出）」であり，テキストから人物に関する属性（生年月日，生誕地，職業，所属組織，職業，出身校，専門，学位，指導教員，国籍など）を抽出するシステムについて記している．個々の属性毎にヒューリスティックな抽出ルールを用いており，表記のばらつきなど各属性に個別的な課題に対処している．作成したシステムで Web ページから人物属性の抽出を競う国際ワークショップに参加し，15 参加システム中で５位の成績を挙げている．

第７章は「Relational Model of Semantic Similarity（意味的類似性の関係モデル）」と題し，これまで２単語間の意味的類似性は各単語の属性の共通性の度合いによって計測されてきたが，２単語間にどのような関係がどれだけ多く存在するか否かでも計測されるという，新しい関係モデルの考え方を提示している．そして，実際に人手で類似度が定められた Miller-Charles の 30 単語対ベンチマーク・データ，及び 353 単語対を含む Word Similarity-353 データセットを実験対象にして，第３章と同様な検索エンジン出力スニペットの周辺文脈単語パターンのクラスタリングによって形成する特徴量と類似単語の対からのその重みの学習，及び特徴量間の相関も考慮することにより，２単語の類似度を計測し，その効果を示している．実験により，この新しい関係モデルによる意味的類似性の計測は，属性類似性による手法と同等な効果を達成できることを示している．

第８章「Conclusions and Future Work（結論と今後の研究）」では，本論文の成果を纏めると共に，類似性についての考察，今後の研究について記している．

以上を要するに，本論文は単語間の属性類似性及び単語対間の関係類似性を Web 検索により計測する効果的な手法を創案，開発し，その性能を実験を通して評価し，既存手法に対する優位性を実証している．単語対に関する検索エンジン出力スニペットに現れる多数の周辺文脈単語パターンを効果的に処理し，ローカルな特徴量として利用する点が共通する大きな特長になっている．関係する課題として，Web における同姓同名人物判別の新手法，別名検出の新手法，人物の属性抽出法も提示し，その性能を実証している．これら手法における創意が優れると共に，手法の実現と適切なデータを用いた性能の実証法も堅固なものであり，Web や人工知能分野で最高峰の国際会議に複数回論文採録されるなど，国際的にも高く評価される研究成果となっている．これらの研究成果により，本論文は電子情報学上貢献するところが大である．

よって本論文は博士（情報理工学）の学位論文として合格と認められる．