審 査 の 結 果 の 要 旨

氏名　　顔　玉蘭

本論文は「Relation Extraction from Web Contents with Linguistic and Web Features（言語分析及び Web 上の情報を用いたコンテンツからの関係の抽出)」と題し, 英文 7 章から成る.

第 1 章「Introduction（序論)」では，本研究の動機と主要な内容を述べている．第一の内容は，自然言語テキストで表される意味概念をコンピュータにも意味を把握できる形の構造的表現に変換する手法に関する研究である．第二の内容は，Web のような大規模テキストコーパスから関係情報を抽出する手法に関する研究である．

第 2 章「Related Work on Relational Extraction（関係抽出についての関連研究)」では，第一，第二の研究課題に関する以下のような共通する関連研究について述べている．即ち，テキストの構文及び依存解析法，教師付き学習法と半数師付き学習法，教師なし学習法（クラスタリング法）についてまとめると共に，Web を代表とする大規模コーパスの記事の性質と，それらからの関係抽出の研究例について述べている．

第 3 章は「A New Shallow Semantic Parser for Describing the Concept Structure of Text（テキスト概念構造記述のための新しい表層的意味パーサ)」であり，第一の研究内容について記している．これは今後とも増大を続ける Web を中心とする情報を，コンピュータがその意味を理解して処理できるようにするのに必要な共通的基盤技術となるものである．ここで，テキストが表す意味概念は，意味役割(semantic role)を中心にエンティティ間を関係付けて構造化して表す．英語に対して使用されている PropBank の意味役割は言語依存性があるのに対し，ここでは多言語翻訳の中間言語（ピボット語）から派生して定められた，言語非依存でユニバーサル性を持つ CDL(Concept Description Language)の関係記述子をテキストから抽出する方法を提示している．テキストの依存構造解析により関係を持つエンティティ対の抽出法，関係の種別を識別するための特徴量を定め，関係が記述されたテキストコーパスを利用し，カーネル関数を用いる教師付き学習により CDL 関係子識別ルールの構成法を示している．出現頻度が低い CDL 関係子に対しては，事例不足で学習が出来ないが，出現頻度が高い CDL 関係子は，テキストからおよそ 87%の精度で認識できることを実験により示している．

第 4 ～ 6 章は，第二の研究内容である Wikipedia テキストからのエンティティ間の意味的関係の抽出法について記している．Wikipedia のテキストを対象とするのは，雑多な Web テキストの中で内容と記述形式の品質が整っているからである．

第 4 章「Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web（Web 情報を用いる Wikipedia テキストマイニングによる教師なし関係抽出)」では，Wikipedia ページタイトルのエンティティとそのページの複数のアンカーエンティティ（別の Wikipedia ページのタイトルエンティティ）との間の関係を抽出する手法を提示している．この関係抽出には二種の特徴を利用する．第一は，エンティティ対を検索エンジンの入力として得られる検索出力のスニペットのエンティティ対周辺の重要語彙（ここでは具体的に重要度を判定した動詞と名詞）を含む語彙パターン集合であり，これを Web 文脈特徴(Web context feature)と呼ぶ．第二は，該当の Wikipedia ページテキストに現れているエンティティ対を含む文の依存解析によって得られる部分木(sub-tree)集合であり，これを言語学的特徴(linguistic feature)と呼ぶ．言語学的特徴は，信頼度は高く雑音的要素は低いものの，データ量は少なく不十分であり，一方，Web 文脈特徴は雑音的要素も含むものの多量のデータを利用できるという相異なる性質を有する. 多数の Web ページから得られるエンティティ対に対するこれら情報に基づき，以下のクラスタリングを行うことにより，エンティティ対間に存在する有意な関係の抽出を行う．このクラスタリングでは，まず上記の重要語彙に基づいてエンティティ対をグループ化してクラスタリングの初期中心を形成し，信頼性が高い言語学的特徴量空間の距離に基づく反復型の K 平均クラスタリングを行うのだが，Web 文脈特徴量空間での Levenshtein 距離（編集距離）で判定してある閾値以上のエンティティ対はクラスタへの併合を行わないようにする．これによって，クラスタに属さない多くのエンティティ対が残ることになる．また，データ量が不十分な言語学的特徴空間での距離に基づくクラスタリングでは，同一の関係を有するエンティティ対であっても別のクラスタになることも多い．そこで，次いで Web 文脈特徴空間での距離の近さに基づいて，クラスタの併合，及び孤立エンティティ対のクラスタの併合を行う．このようにして形成されたエンティティ対のクラスタが同種の関係を有することになる．本手法の評価として，米国 CEO(Chief Executive Officer)と Companies についての Wikipedia ページ（計約 1,000 記事，11,400 エンティティ対）を対象にして実験を行い, CEO である人物とその企業名, 生年, 出身大学, 学位, 結婚相手等の関係, 企業と CEO,

創業者，本社所在材地，創業年，買収企業等の関係が抽出でき，単独の特徴空間を使用するよりも良い結果が得られることを実証し，言語学的特徴が精度向上に役立ち，Web 文脈特徴が被覆率向上に貢献することを示している．

第 5 章「Multi-view Clustering Learning Approach for Relational Extraction from Wikipedia Texts（Wikipedia テキストからの関係抽出のためのマルチビュー・クラスタリング学習のアプローチ)」では，上記の言語学的特徴と Web 文脈特徴を用い，エンティティ対のクラスタリングだけでなく，特徴量のクラスタリングも同時に実行する双対型共クラスタリング(dual co-clustering)の手法を提示している．本手法は，情報量基準による共クラスタリング手法と自己教示(self-taught)クラスタリング手法を基にしているが，両特徴量を用いて新しい媒介項を導入した共クラスタリング手法を実現している．具体的には，まず，エンティティ対の言語学的特徴によるクラスタリングと Web 文脈特徴によるクラスタリングの共通項が多い集合を求め，これを両特徴量クラスタリングの橋渡しの役割を果たす共通クラスタ集合とする．この共通クラスタのエンティティ対を媒介項として，両特徴量の各々のクラスタリングの情報量的目的関数を定め，これを最適化する反復計算により，両特徴量のクラスタリングを求める．両特徴量のクラスタ集合を重み付けで結合して新たな特徴量次元として，エンティティ対をクラスタリングし，有意な関係を有するエンティティ対の抽出を行う．特徴量のクラスタリングは，データの希薄性に対処して安定的な結果をもたらすことに貢献する．本手法の評価として，婚姻，社長，CEO，生年月日等 13 種の関係を持つ Wikipedia ページから得た 3,800 程のエンティティ対を対象にした実験により，既存手法より優れた関係抽出結果が得られることを示している．

第 6 章「Multi-view Bootstrapping Approach by Exploring Web Features and Linguistic Features（Web 特徴と言語学的特徴を探索するマルチビュー・ブートストラッピング・アプローチ)」では，関係種別ラベルが付された少数のエンティティ対事例をシードとして，多数のラベルなしエンティティ対を識別する第 4，5 章に共通する Web 文脈特徴と言語学的特徴を用いる半教師付き学習手法を提示している．ここで，両特徴量に関しては，データ希薄性問題等を回避するために，別途に所属研究室で開発された分布仮説に基づく効率的クラスタリング法により，クラスタリングして利用する．各々の特徴空間で現在利用できる関係ラベル付きエンティティ対の関係種別を識別するルールの学習を行い，このルールにより残りの多数の関係ラベルなしエンティティ対の関係識別を行う．両特徴空間で識別された関係が情報量的に整合しているとされたエンティティ対のみを新たな関係ラベル付きエンティティ対集合に追加し，特徴量の新たなクラスタリング，関係識別ルールの学習を反復する．即ち，両特徴空間での識別が一致しない不確かとみなされるデータを排除しつつ，ブースティングによる半教師付き学習であるクラスタリングを実現し，これによって多数のラベルなしデータも活用する優位な関係を有するエンティティ対の抽出を実現している．本手法の評価として，Wikipedia テキストから得た人物に関する職位，生年，出身大学，死亡年の 4 種関係をもつ約 580 エンティティ対を対象にした実験を行い，両特徴量空間を用い，不確かなデータを排除しつつ進めるマルチビュー・ブートストラッピングが有効であり，既存手法より優れた関係抽出結果が得られることを実証している．

第 7 章は「Discussion and Future Work（議論と今後の研究)」であり，本論文の成果をまとめていると共に，その新規性と今後の研究方向に言及している．

以上を要するに，本論文は Web からの関係情報抽出に関し，以下の研究成果を記している．第一は，テキストが表す意味概念をコンピュータにも意味を把握できるような形の構造的表現に変換するため，テキスト依存構造解析結果より有効な素性特徴を定め，カーネル関数を用いる教師付き学習によりテキストに現れるエンティティ間の関係を識別するルールの学習法を示している．第二は Web を代表とする大規模テキストコーパスからエンティティ対間に存在する有意な関係を抽出する課題に関し，検索エンジン出力のスニペットから得られる Web 文脈特徴（エンティティ対周辺の語彙パターン集合）と具体的な文の依存解析によって得られる言語学的特徴（部分木の集合）を相補的に利用してのクラスタリングによる有意関係抽出法，同様に両特徴量を用いてエンティティ対と特徴量のクラスタリングを同時に実行する双対型共クラスタリング法による有意関係抽出法，同様に両特徴量を用い関係種別レベルが付された少数のエンティティ対事例をシードとして多数のラベルなしエンティティ対を識別する，ブートストラッピングによる半教師付き学習のクラスタリングによる有意関係抽出法を考案，実現している．これら手法の有効性を Wikipedia テキストを用いての実験により実証しおり，電子情報学上貢献するところが少なくない．

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．