審 査 の 結 果 の 要 旨

氏名　グェン トアン ドゥク (Nguyen Tuan Duc)

本論文は「Latent Relational Web Search Engine Based on the Relational Similarity between Entity Pairs (エンティティペア間の関係類似性を利用するウェブ潜在関係検索エンジン)」と題し，英文で記されており，７章から成る．

第１章「Introduction(序章)」では，まず潜在関係検索とは Web 検索において，例えば{(Japan, Mt. Fuji), (Germany, ?)} と問い合わせると，Germany で一番高い山として “ ? = Zugspitze” を答える新タイプの検索エンジンであることを説明している．一般的に記すとクエリ（問合せ）{(A, B), (C, ?)} において(A, B)をソースペア，?を含む(C, ?)をターゲットペアと称し，(A, B)間に成立するのと同様，或いは類似関係を有する(C, ?) の “?” を求める検索である．（ここで，“?” は最後の位置だけでなく上記の A,B,C の位置のどこに置いても良い．）更に，{(日本，富士山),(Germany, ?)}のように，ソースペアとターゲットペアは異なる言語であっても可能な言語横断型(cross-lingual)関係検索を実現したことを述べている．そして，このような研究を行った背景と動機を述べている．

第２章は「Relational Similarity and Latent Relational Research(関係類似性と潜在関係検索)」であり，最初に本研究で基礎としたエンティティペア間の類似性計測法について説明している．これは Web テキスト中におけるエンティティペア間の周辺語彙系列パターンの分布の類似性が高いペアは，関係類似性も高いと見なせる分布仮説(Distributional Hypothesis) に基づくものである．周辺語彙系列パターンは２つのエンティティ(単語)を含む文から，エンティティペア間だけでなく，その前後も含めて所定の閾値以上存在するパターンを選ぶ．ターゲットペア(C, ?) については，C の近傍で共起するエンティティ(単語)とのペアについて周辺語彙系列パターンを求め，ソースペア(A, B)の周辺語彙系列パターンとの類似性からランキングする．これは既存研究を参照したものだが，実用的な時間で検索結果を出力する高速検索を実現するためには事前処理によるインデックス作成が不可欠となり，この部分が本研究のオリジナルな成果となっている．また，言語横断型潜在関係検索も実現可能にしていることも，本研究のオリジナルな成果となっている．

第３章は「Related Work(関連研究)」であり，これまでの構造写像理論(Structure Mapping Theory)などの類推，特異値分解を用いる潜在関係写像エンジン(Latent Relational Mapping Engine)，高レベル認知・類推のモデルなどの関連研究について記している．また，エンティティペア間の関係類似性の，本研究で用いた計測法以外の計測法について言及している．その後に，Web コーパスからの下位語(hyponyms)や上位語(hypernym)抽出法，部分全体の関係の語(meronym)抽出法などの既存研究について紹介している．

既存の関係検索システムに関係する研究についてもまとめている．それらは，“Muslim church”に対し “mosque”を答え，“Greek A”に対して “α”を答えるシステム，スロット充填システムとして “ * is the president of France.” の問合せに対してフランス大統領名を答えるシステム，INDUCES(asprin, ?) の問い合わせに対してアスピリンが誘発する効果を答えるシステム等を紹介している．本研究の関係検索と同様なシステムもほぼ同時期に開発されたものが報告されているが，これはエンティティペア間の限られた語彙系列パターンを利用する方法であり，本研究の方法はより包括的であり，かつインデックス化による高精度，高速化を実現していると述べている．言語横断型検索に関して本研究は，翻訳辞書にはしばしば存在しない固有名詞エンティティを検索対象にすることから，既存の単なる語の翻訳に頼るのとは異なる方法を採るとしている．

第４章「Retrieval Model for Monolingual Latent Relational Search(単一言語の潜在関係の検索モデル)」では，上記した本研究の潜在関係検索法について詳述している．即ち，エンティティの抽出は所定以下の近さで１文中に共起する固有名詞ペアをエンティティペアとして，両エンティティ間と前後３単語を含む一定以上の出現頻度を持つ周辺語彙系列パターン(*(wild card)を含むパターンも含む)を抽出する．この際，変動を抑制するため，単語は語尾変化を除き語幹にする(stemming)．このようにして得られる周辺語彙系列パターンには付随するエンテ

ィティペアが記録されるのだが，同様なエンティティペアを有する周辺語彙系列パターンは意味的に類似な関係を表していると考えることができるので，これらをクラスタ化する．例えば，買収関係を表す“X acquired Y.”と“X bought Y.”は多くのエンティティペアを共有するので，意味的に近いと判断でき，一つの項目にクラスタ化する．高速検索を可能とするために，各エンティティペアに対する複数クラスタ化語彙系列パターン（出現回数付き）のインデックスファイル，その転置インデックスに相当する各クラスタ化語彙系列パターンに対応する複数エンティティペア（出現回数付き）のインデックスファイルを作成し，保持する．エンティティペア間の関係類似性は，基本的にコサイン類似度により計算している．

第５章は「Retrieval Model for Cross-Lingual Latent Relational Search(言語横断型潜在関係検索のための検索モデル)」であり，英語－日本語を具体例として言語横断型関係検索の実現法を記している．両言語テキストからのエンティティペア抽出，語彙系列パターンの抽出までは第４章の場合と同様である．両言語の橋渡しは，例えば日本語テキスト中には(グーグル，ユーチューブ)ペアと(Google，YouTube)が共に使われて出現することの利用，及び周辺語彙系列パターンを機械翻訳して他方の言語の語彙パターンにして利用（この翻訳は*を含まないパターンに限っている）によって実現している．そして 2 段階のクラスタ化により，両言語を統合した周辺語彙系列パターンのクラスタ化を実現している．言語横断型の場合，エンティティペアもクラスタ化することにより，(グーグル，ユーチューブ)と(Google，YouTube)等を同一のクラスタ化項目にまとめることができる．このように言語横断型の場合もほぼ同様にインデックスファイル，転置インデックスファイルを作成でき，潜在関係検索が実現できることを記している．そして，本手法による言語横断型潜在関係検索は，実験によりベースラインとなる既存手法よりも優れた性能を達成することを示している．

第６章は「Milresh: A Large-Scale Latent Relational Search Engine based on the Proposed Model (Milresh : 提案モデルに基づく大規模潜在関係検索エンジン)」であり，実装した Milresh システムの構成と key-value store として実装した大規模インデックス構成を記し，性能評価について述べている．このシステムはクラウドコンピュータ環境で実装しており，Hadoop など並列分散プログラミングを活用している．検索結果はランキング付きで出力され，併せて根拠となったセンテンスも出力される．判定の基になるデータ規模は、700 万 Wikipedia ページ（内 160 万が日本語ページ）から得た 2 億 1300 万センテンスを処理し，668 万エンティティ，3,077 万エンティティペア，9 億 4585 万周辺語彙系列パターンを抽出し，潜在関係検索のためのインデックスを作成している．このインデックス作成の前処理に要した時間は，各々6CPU を持つ 5 並列マシンを用いて 7 日である．固有名詞を答える質問応答用データセットを用い，等価な関係検索を行う評価実験を行い，英語単一言語クエリにおいては 43%程，英語－日本語横断型クエリにおいては 34%程の正解を出力できることを実証している．その他の多くのクエリに対する検索性能も示している．これらの検索時間は 3 秒程度であり，事前のインデックス化により実用的な時間で検索が実現できることを実証している．

第７章「Conclusion(結論)」では，本論文の成果をまとめ，残された課題に言及している．

以上を要するに，本論文は{(Tokyo, Japan), (?, France)}といった問い合わせに対し，“?=Paris”と答えるような新タイプの潜在関係検索エンジンを，テキスト中のエンティティペアに対する周辺語彙系列パターンの分布の類似性計算に基づく検索を原理とし，表記ゆらぎを吸収することによる高精度化と実用的な検索時間を達成するために，周辺語彙系列パターンのクラスタ化を含むインデックスを構成することにより実現する方法を考案している．更に，英語－日本語といったような言語横断型潜在関係検索の実現法も考案している．そして，具体的に潜在関係検索エンジンを実装し，大規模テキストコーパスからインデックスを作成し，性能を評価，実証している．このように新機能 Web 情報検索エンジンを実用に近い形で開発したことで情報理工学に貢献し，創造的実践の観点から価値が認められる．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる．