審査の結果の要旨

氏名　　楊　潔（Yang Jie）

本論文は「Collective Semantic Annotation for Web Text: Triple Tagging and Triple Extraction（Web テキストへの集団による意味的アノテーション：トリプルタギングとトリプル抽出）と題し，英文で記されており 9 章から成る．

第 1 章は「Introduction（序論）」であり，まず Web 検索の精度を高めるなどの高度化を図るには，Web ページにその意味的内容を表すアノテーションを付すことが効果的となるが，その方法には，1）オントロジー・セントリックな意味的アノテーション（オントロジーとは上位下位関係等が定められた基本語彙大系）と，2）ユーザ・セントリックな意味的アノテーションがあるとしている．1）はセマンティック Web で採られているアプローチであり，広汎な領域で広く認められる統制されたオントロジーを確立することが難しく，かつ一般人にはオントロジーに従ってアノテーションを与えることは敷居が高く，なかなか広まらないという問題を抱えている．これに対し 2）は，社会的タギング，集団的アノテーション，フォークソノミーなどの形で普及しているが，アノテーションとして複数ユーザが自由に付与するキーワードから的確に意味を把握することが課題となる．本論文は 2）のアプローチを採り，次の論点に対して答えようとして行った研究であることを述べている．

・如何にして表現力が有り，ユーザが理解容易で，かつ，コンピュータが意味を把握できる社会的集団によるアノテーションの形式を見出すか．
・如何にしてアノテーション・タスクを自動化するか．

そして，本論文は具体的に 3 つ組構造データ（典型的には，主語，述語，目的語から成る）を用いるトリプルタギングによるシステムを提案し，設計方針，自然言語テキストからのトリプル（3 つ組構造データ）の抽出法，評価用システム作成等について記しているとしている．

第 2 章は「Review of Annotation System（アノテーションシステム概説）」である．アノテーションシステムを特徴付ける 8 次元の軸（標準データフォーマット，アノテーションの蓄積形式，アノテーションの粒度，アノテーションのプリミティブ，アノテーション語彙のソース，ユーザ中心のデザイン法，アノテーションの消費者，アノテーションによるサービス）について述べ，これに基づいて既存の各種システムの位置付けを明らかにしている．また，オントロジーに基づくアプローチ，社会集団的アプローチ，両者の橋渡し的アプローチに大別し，それぞれの代表的システムについて説明している．そして，本論文のシステムの位置付けは社会集団的アプローチに立脚するが，アノテーションのプリミティブとしてトリプルタグを用いることで拡張を図ったものであると述べている．

第 3 章「Triple Tagging Model（トリプルタギング・モデル）」では，本論文で提案するアノテーションシステムは典型的には（主語，述語，目的語）をとるトリプルをプリミティブ・データに用いるものであるが，システムを構成する上で必要となる他の要素も含めた基礎となるモデルについて述べている．トリプルは標準化されている RDF(Resource Description Format)に従って記述，管理される．また，トリプルタグを作成したユーザ名，作成日時も記録され，管理されるトリプルタグの集合から相互に関連付けられたタグ・グラフが構成され，グラフ照合による検索，及び情報の視覚化やナビゲーションに用いられる．そして，トリプルタグを記述するガイドラインを示している．

第 4 章「Model Evaluation（モデルの評価）」では，提案のトリプルタギングを第 2 章に記した 8 次元の軸の観点から評価するために，ユーザからのデータの収集と分析を行うケーススタディを行い，既存のアノテーションシステムに対する利点を示している．

第 5 章は「Sentence-Based Triple Extraction（センテンスに基づくトリプル抽出）」であり，ユー

ザのトリプル作成を促進する自然言語処理に基づくトリプルの自動抽出法について検討している．これは文章からの２項関係抽出であり，まず最初に関連研究についてまとめており，特に本研究に利用する核技術としてのトリー・カーネル関数によるセンテンス・トリー類似度に焦点を当ててまとめている．そして，本研究のトリプル自動抽出のための，依存解析の基づくセンテンストリーの類似度計算を各ノードの重要性を考慮して精度を高める，ペナルティ付きトリー類似度を導入している．

第６章「An Interactive Approach for Triple Extraction（トリプル抽出のための対話的アプローチ)」では，ユーザが内容を表す上で重要と考え選択したセンテンスからトリプルタグを自動抽出するに際し，これまでにトリプルタグ化されているセンテンスとのセンテンス類似度計算に基づくkNN（k Nearest Neighbor, k 最近傍）法により抽出すべきトリプル候補を見出し，有効であるとして抽出するトリプルを決定する方法を示している．具体的には以下の手順を採っている．入力センテンス（英文）を依存解析して依存木（トリー）を作成する．依存関係にある主語と目的語の候補ペア語を見出し，このペア語に関する最小木を作成し，これを第５章に記したペナルティ付きトリー類似度によって既存トリプルタグ・データと比較して k 最近傍データを求め，類似度の数値によりフィルタリングして抽出するトリプルを決定する．この処理過程では，WordNet（英単語の上位下位関係等を規定した辞書）に基づく単語間の類似度，形態素解析結果，名辞エンティティ分類結果の情報も利用している．

第７章「Experiments（実験)」では，第６章の方法を実験システムにより検証している．まず，協力者の助力を得て，Wikipedia の音楽バンドの Web ページよりトリプルタグを人手で作成し（363 センテンス，774 トリプルタグ)，これを実験用トリプルタグ・コーパスにしている．そして，これには含まれない音楽バンドの Wikipedia のページの注目されるセンテンスに関し，人手で抽出したトリプルタグと上記コーパスを用いて第６章の方法で得たトリプルタグを比較し，43%程の精度，56%程の再現率でトリプルタグの自動抽出が可能となることを示している．また，ペナルティ付きトリー類似度の有効性も実証している．

第８章「Implementation System: Triple-Note（実装システム：Triple-Note)」では，グラフィカル・ユーザインタフェースも含めて実装したトリプルタギングシステム：Triple-Note について記している．Web ページの内容をよく表しているセンテンスを選択すると，第５，６章の方法に基づいて有効と考えられるトリプルタグが自動抽出され，ユーザはこの中から適当と考えるトリプルタグを選択し，登録する形式をとる．トリプルタグ・グラフの表示は関係する項目の一覧を可能にしている．トリプルタグの１項目あるいは２項目を＊にすることにより，この＊にマッチするトリプルタグ及びそれに付随する Web ページを検索する機能も与えている．

第９章「Conclusions and Future Research Directions（結論と今後の研究方向)」では，本論文の成果をまとめると共に，今後の研究方向に言及している．

以上のように，本研究は Web 検索・利用の高度化を可能にするために，社会的集団によりアノテーションとして３つ組構造データを Web ページに付与するトリプルタギングに関して，自然言語文からトリプルタグ候補を自動抽出する方法等の考案・開発と共に，効果的実現に必要なシステム的検討，並びにシステムの実装による効果の実証を行っており，この分野で世界的に認められるレベルの貢献を果している．すなわち，本研究は情報理工学に関する研究的意義と共に，情報理工学における創造的実践に関し高い価値が認められる．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．