論文内容の要旨

博士論文題目　英文契約書の電子文書化に関する研究

申請者　相良 かおる

（論文内容の要旨）

本論文では，英文契約書の構造文書化と条文の統語解析に関して，次の項目について述べている．(1) 構造文書化に必要な専門用語などの辞書類の作成．(2) 電子化における索引語の作成ならびに情報検索の支援を目的とした，(a) 語の基本形，(b) 条項の内容を表すための重要語，(c) 語の類似度計算とクラスタリング．(3) 契約書の構造文書化に必要な条文の解析方法．個々の内容は次のようである．

(1) 英文契約書の文例集および実例から，品詞辞書，用語辞書，連結語句パターンなどを作成している．

(2) これを基にして，次の概念を定義し，契約文例からそれらを抽出している．

(a)「語の基本形」は，対象とする語から語形を変化させる屈折接辞を取り除いた「語幹」から更に語の品詞を変える派生接辞を取り除いたものである．「語の基本形」を求めるために 97 種類の接尾辞から品詞を変える接尾辞 85 種類を抽出して，語形を変換させる 262 種類の規則を作成している．それによって，(1) で作成した用語辞書に含まれる 5,804 種の語を 2,854 種の基本形にまとめている．

(b) 契約文書では条項ごとに内容がまとめられているという特徴に着目して，索引語を抽出する際に使われる TF.IDF 法を応用して，技術取引に関する契約書式集から，34 条項を対象に，重要語を求めている．それを用いて，異なる文例において条項名の推定を行い，書式集から抽出したテストデータで 79 %の正しさで条項名の推定ができ，かつ「語の基本形」を導入することで，平均 85 %の正解を得ることを確認している．

(c) 類義語は文書の解析や情報検索を行う際に重要なデータとなる．そのため，類義語辞書を作成する際の作業量を減らすように，関連度を用いた類似度計算の提案を行っている．「関連度」とは語句間の係り受けの強さを表す尺度で，条文を解析する際，妥当な係り受けを決定するために提案したものである．次に，(名詞，動詞)，(動詞，名詞)，(形容詞，名詞)，(前置詞，名詞) の 2 つ組の各々の関連度を用いて名詞間の類似度を求めて，ファジイ同値関係から，名詞の全体集合を同値類に分類している．用語辞書の 894 種の名詞について，283 種の名詞からなる 99 個の同値類を求め，意味を確認して 84 種の名詞からなる 34 個の類義クラスを求めている．

(3) 長文で複雑な契約書の条文を完全に統語解析することは困難である．本研究では，条文から統語構造を抽出する手法を提案して，実装した結果を述べている．ここで「統語構造の抽出」とは，一つの解析木を求める統語解析とは異なり，(1) 修飾－被修飾関係，および (2) 主部・述部を独立に抽出することである．本手法の特徴は，動詞を中心とするパターン情報を用いて解析する方法と，文法規則による解析の二つの手法を併用して解析を行っている点である．

氏　名	相良 かおる

(論文審査結果の要旨)

本論文は，英文契約文書の電子文書化に関する研究について述べたものである．

(1) 「語の基本形」の提案：「語の基本形」を定義して，「基の語」,「語幹」,「語の基本形」のそれぞれから重要語を抽出した．3 種の重要語を使った条項名の推定実験を行い，「語の基本形」は「基の語」の 53 %のデータ量で 1.06 倍の正解率を得ている．少ないデータ量で同義の多様なテキストデータに柔軟に対応できること，重要語の抽出に有効な効果を発揮していることを示した．

(2) 重要語の定義と重要語の抽出：技術取引に関する契約書式集に含まれる 34 種類の条項における重要語の一覧を作成し，テスト用の条文に含まれる条項名の推定実験を行った．その結果,「語の基本形」を用いて 85 %の正しさで推定できたことを示した．本研究で作成した重要語が，国際取引における契約文書の条項の内容の推定に，また情報検索する際のキーワードとして有用であるとしている．

(3) 共起する語の二つ組の関連度：文書の解析に当たって，文法上，複数の係り受けが考えられる場合に妥当な係り受けを決定するために関連度を導入している．関連度は，語の二つ組の同時出現回数に，各単語の汎用性の相違を重みとして掛けたものである．その中で，とくに，動詞と名詞の関連度について，他の方法を比較し，提案した方法が有効であることを示した．

(4) 語間の類似度計算とファジィ関係行列を用いたクラスタリングの提案：類義語辞書を作成する際の作業量を軽減するために，関連度を用いた類似度計算の提案を行っている．関連度をベクトルとした内積法によって名詞間の類似度を求めている．894 種の名詞の類似度を基に 894 × 894 類似関係行列を作成し，ファジイ同値関係を導き，99 個の同値類を求め，そこから 34 種の類義クラスを作成している．

(5) 契約文書における統語構造の抽出：文書に含まれる構成語句の統語構造，語義，参照関係などを構造化するために，文法規則と動詞を中心とした抽出パターンとを併用して処理を行う統語解析の手法の提案している．それは，等位関係，相関接続の関係など 7 種類の統語構造の抽出を再帰的に行う方法である．秘密保持条項に出現する 57 種の動詞に関する 143 の抽出パターンから，条文の長さに関係なく結果が得られること，出力した結果を見直し修正ができること，解析の精度を上げるために緒規則・データを抽出できることなどの利点を示している．

本研究は，英文契約書を対象にして，効率の良い情報検索および内容の特徴抽出方式を確立し，契約文書の電子文書化 (構造化) に対する基盤を構築している．提案方式の多くの部分は一般の英文書にも適用できる汎用性を備えており，学術上，応用上寄与するところが多い．よって，本論文は博士 (工学) の学位論文として価値あるものと認める．