論文の内容の要旨

論文題目　　多言語シソーラス自動構築手法開発の為の実験的研究

氏　　名　　辻　慶太

本論文では，対訳コーパスに基づいて，多言語シソーラスを効率的に自動構築する手法を開発・提案した。ここで多言語シソーラスとは「複数言語における同じ意味の語をクラスタ化した語彙集」である。

第１章では，本研究の動機・背景について説明した。現代社会においては，インターネット等を通して様々な言語で情報が発信・蓄積されているが，多言語の情報を適切に取捨選択するには，それに対応した効率的な情報検索システムの実現が必要である。複数言語における同じ意味の語を明示した多言語シソーラスは，その実現を可能にする。また様々な言語の情報が流通するにつれ，今後人々は母語以外の言語を学習する必要に迫られていく。多言語シソーラスは，そうした学習用の辞書としても有効に機能する。このように多言語シソーラスには多くの有用性が認められるが，シソーラスを手作業で構築・更新するには，かなりの手間と時間がかかる。そのため，シソーラスは最近に生まれた語を収録していないという状況が生じやすい。情報検索システムや学習の場において，新語に対するニーズは高いと考えられるので，新たに使われるようになった語を遅滞なく多言語シソーラスに取り込めるような自動構築手法の開発は，きわめて重要かつ実践的な研究課題である。こうした見地から本論文では，新語の適切な扱いに特に焦点を当て，対訳コーパスから多言語シソーラスを自動構築する手法の開発に取り組んだ。新語はコーパス中の出現頻度が低いので，低頻度語を適切に扱う方法を構築できれば，新語の処理方法としては十分である。そこで本論文では新語の適切な扱いを目標とはしているが，新語を含むより広い低頻度語の適切な抽出・関係付け方法を追求した。

対象言語は日本語と英語とした。対訳コーパスとしては，国立情報学研究所の学会発表

データベースに含まれる論文の日英抄録対及び日英タイトル対を用いた。これらデータベースには，常に新たな論文の抄録・タイトルが追加されており，新語が多く含まれている。対象分野は，人工知能，林学，情報処理，建築学の４分野とし，それぞれの分野に関して，1,000 抄録対，1,000 タイトル対を用いた。情報処理分野と建築学分野に関しては，9,000 抄録対，9,000 タイトル対も用いた。従って，全部で 12 種類の対訳コーパスを実験対象とした。抽出対象は，新語や情報検索の現状を考え，名詞性の単位語とした。

第２章では，関連する先行研究について概観した。近年，対訳コーパスの増加を受けて，それに基づく多言語シソーラス自動構築研究が，活発に行われるようになっている。だが従来のシソーラス自動構築研究は，どの言語においても通用する普遍的な構築手法を模索する傾向があり，対訳コーパス中の語の頻度に基づく統計的な手法に偏る面があった。そのためコーパス中の頻度が本来的に低い語である新語を，適切に扱うことができない手法が数多く提案されてきた。あるいは新語の扱いが本質的に重要であることが認識されていなかったとも言える。それに対して本論文は，新語の重要性を認め，また頻度だけでなく，日本語と英語という言語対固有の性質を利用することで，新語を適切に扱える手法を開発したものである。さらに先行研究では，訳語対とは言えない対に対して，非常に小さい対訳確率値を付与する場合が散見されるが，本論文では，多言語シソーラスにおいて重要なのは意味が同じか否かという語彙的な関係であるという立場を明確にした。

第３章では，本論文が前提とするいくつかの側面に関して，基礎的調査を行い，それらの結果を踏まえて独自の多言語シソーラス自動構築手法を提案した。まず訳語対辞書である EDICT を，既存の多言語シソーラスの１つとみなし，この EDICT との照合を通して，上記対訳コーパス中の低頻度訳語対の多くは，既存の多言語シソーラスに収録されていないことを明らかにした。一方で低頻度訳語対の一部は，一過性の対として消えたりはせず，その意味で重要な抽出対象であることを時系列的分析によって明らかにした。

低頻度訳語対の多くは，頻度に基づく従来の手法では抽出できない。このことを実証するため，対訳コーパスでのセグメント（訳語対を抽出する範囲）における「内部完全共起」と「外部完全共起」の概念を提唱した。まず訳語対(X, Y)があり，語 A が，語 X の現れるセグメントに必ず現れ，語 X の現れないセグメントには現れない場合，語 A は訳語対(X, Y)と完全共起していると定義した。さらに語 A が語 X と同じ言語に属する場合，語 A は訳語対(X, Y)と内部完全共起していると定義し，語 A が語 X と異なる言語に属する場合は，語 A は訳語対(X, Y)と外部完全共起していると定義した。内部完全共起が起きている場合，従来の頻度に基づく抽出手法は，語 Y の訳語が語 A なのか語 X なのか決定することが出来ない。また外部完全共起が起きている場合，従来の頻度に基づく抽出手法は，語 A・語 X は，語 X・語 Y と同等あるいはそれ以上に訳語対である可能性が高いと判断してしまう。本論文では，対訳コーパスにおける低頻度訳語対の多くが，他の語に完全共起されており，頻度に基づく手法では適切に抽出し得ないことを実際の調査に基づいて示した。

日英の低頻度訳語対には，借用語と元の語という借用語系の訳語対が多い。借用語系訳語対は，語の頻度情報を利用することなく，一定の翻字パターンに基づいて，対訳コーパスから自動抽出することが出来る。さらにそうして低頻度訳語対の一部を抽出・除外すると，残りの低頻度訳語対における完全共起が，ある程度解消される。例えばセグメントにおいて，語 A が他の語 B などと借用語系訳語対になっていた場合，対(A, B)が抽出・除外

されることで，セグメントには対(X, Y)のみが残り，完全共起は解消される。第３章では，対訳コーパス中に借用語系訳語対がどの程度存在するか，またそれらが抽出・除外された場合，どの程度完全共起が解消されるかを実際の調査に基づいて明らかにした。

上記の結果を踏まえて，シソーラス自動構築手法として以下の３ステップから成る手法を提案した。まず，(1)翻字に基づいて，借用語系訳語対を抽出・除外する，(2)残った候補語から，頻度に基づく手法で残りの訳語対を抽出する，(3)抽出した全訳語対を語を頂点とする辺とみなしてグラフを構築し，グラフ理論に基づく手法で，同じ意味の語をクラスタ化する，という手法である。

第４章では，上記提案手法がどの程度有効であるかを，実際の多言語シソーラス自動構築実験に基づいて示した。以下では，上記(1)(2)(3)の各ステップについて説明する。

(1)に関する提案手法では，まず次の３点を仮定した。即ち，(a)日本語が英語を借用する場合，語は日本語の拍に基づいて翻字され，それらの対応は安定的である，(b)これらの翻字は前後の音・拍から独立して行われる，(c)翻字パターンは時間経過や分野の違いによって大きく変化することはない，の３点である。本論文で提案した手法は，まず訳語対辞書から抽出した翻字パターンに基づいて，日本語候補語を拍単位で英文字列に変換し，それら英文字列と英語候補語との類似度を Dice 係数的な尺度で測定し，その値が高い対を訳語対と判断するというものである。実験の結果，本手法は先行研究手法よりも高い精度・再現率で借用語系訳語対が抽出できること，また上記の翻字パターンは，単純なヘボン式翻字パターンより有効であり，かつ上記尺度は他のいくつかの尺度候補より有効であることを確認した。

(2)の頻度に基づく抽出手法は，これまで精力的に研究され，非常に精緻な手法が提案されており，そうした手法によって，他語と完全共起していない訳語対は，比較的高いパフォーマンスで抽出することが出来る。本論文では頻度に基づく代表的抽出手法である Melamed の手法と Hiemstra の手法を改良し，その改良手法が，先行研究手法のすべてを上回るパフォーマンスを持つことを実証した。主な改良点は，最終的な訳語可能性値に周辺頻度を組み込んだ点である。周辺頻度を取り込んだ尺度の方が，取り込まない尺度より抽出結果が良いというのは，一般性のある知見と思われる。また本論文では，頻度に基づく手法で従来よく用いられてきた一般相互情報量は，大きな弱点を持つことを証明した。

さらに本論文では，実際のデータに基づいて(1)(2)の組み合わせが有効であり，両者を組み合わせた手法が，翻字に基づく手法単独，頻度に基づく手法単独よりも，明らかにパフォーマンスが高いことを示した。そうした傾向は，一方の語の頻度が１で，その意味で非常に完全共起されやすい訳語対に関する抽出結果でも観察された。低頻度訳語対の自動抽出に関するこれらの成果は，本論文の主要なオリジナルな点と言えよう。

(3)に関しては相澤・影浦の手法を改良し，低頻度クラスタの抽出において，彼らのパフォーマンスを明らかに上回る手法を提案した。主な改良点は，グラフの辺の重みとして，語の共起頻度そのものではなく，訳語である確度を採用した点である。

本論文では，従来あまり注目されて来なかった新語に焦点を当て，言語固有の性質を利用した，従来にない多言語シソーラス自動構築手法を開発した。本論文により，シソーラス自動構築のために，対訳コーパスを現実的に活用する要件，そして現時点での１つの最適解を示すことができたと考える。