

Title	主成分分析を用いた未登録語のシソーラスへの追加
Author(s)	鈴木, 勝仁
Citation	
Issue Date	1998-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1128
Rights	
Description	Supervisor:奥村 学, 情報科学研究科, 修士

主成分分析を用いた未登録語のシソーラスへの追加

鈴木勝仁

北陸先端科学技術大学院大学 情報科学研究科

1997 年 2 月 13 日

キーワード： 主成分分析, スパースネスの問題, 多義語の問題, 共起データ.

単語間の上位下位関係や同義関係を記述したシソーラスは、自然言語処理分野における最も重要な知識源の 1 つである。シソーラスは、多義性の解消や類義語の獲得など、多くの研究で利用されている。これまでのシソーラスは、手作業で構築され、かなり広範囲な語彙が登録されている。しかし、特定領域で扱われる自然言語処理システムの立場からすると、シソーラスは、不必要となる一般的な単語を多く含み、必要となる領域特有の単語を含んでいないため、語彙が不十分であると考えられる。また、シソーラスの構築後は、新たに単語を加えることはなく、語彙の変化に対応できないなどの問題がある。これら問題を解決するために、本研究では、シソーラスに存在しない単語（未登録語）の配置先を推定する方法を提案する。

未登録語をシソーラスへ配置する方法として、文字情報に基づくものがある。漢字は、表意文字であり、それは語の意味に関わる有用な情報になる。日本語シソーラスを対象とするとき、文字（漢字）の情報を使うことにより、比較的容易に意味的な情報を得ることができる。例えば、「自然言語」は「言語」と同じノードに、「経済問題」は「問題」と同じノードに属す。しかし、この文字情報は、未登録語がカタカナや簡略化された語の場合には適用できないため、有効ではない。

他の方法として、コーパスから得た共起データに基づくものがある。例えば、格関係にある名詞と動詞の共起データによって、シソーラス中の名詞と未登録語を動詞要素（格と動詞の組）で表現でき、それら名詞と未登録語の動詞要素の似かより具合で、配置先を決めることができる。共起データに基づく研究には、浦本と、徳永、中山のがある。浦本の研究では、人がシソーラスを構築したときに使われた情報（分類基準）が記されていないために、その分類基準を共起データから抽出し、その分類基準とコサイン距離を用いて、未登録語を ISAMAP という上位下位シソーラスに配置する方法を提案した。浦本の方法では、未登録語は幾つか隣接したノード集合の中に配置される。また、徳永の研究では、未登録語

がシソーラスのノードに属す確率を推定するために共起データを用いる。未登録語は、その確率を基にして分類語彙表という分類シソーラスのノードに配置される。徳永の方法では、未登録語に幾つかの配置先を提示する。そして、中山の研究では、浦本とは異なる方法で、分類基準を共起データから抽出し、その基準を用いて、未登録語を分類語彙表に配置する方法を提案した。中山の方法でも、未登録語に幾つかの配置先を提示する。共起データを用いることで、単語間の類似度計算が可能になり、未登録語の配置先が推定できる。しかし、共起データを用いて類似度を計算する場合、以下の2つの基本的な問題がある。

スパースネスの問題：コーパスから得られる共起データは不完全であることや、動詞概念の表現方法は普通幾つもあるために、未登録語の動詞要素と、配置先の単語の動詞要素とが一致しない可能性がある。

多義語の問題：多くの単語¹は、複数の意味があるため、未登録語の動詞要素が、配置先と関係のない単語の動詞要素と一致する可能性がある。

スパースネスの問題において、浦本の研究では、ノードに直接属す単語と未登録語との類似度を計算するため、スパースネスの問題が生じやすいと考えられる。徳永の研究では、未登録語と単語の類似度ではなく、未登録語と単語集合との類似度を計算するため、共起データの不足分を補うことができるが、従来のスパースネス解消方法と比べると弱い。そのスパースネス解消方法とは、シソーラスの構造を用いて、意味的に似た動詞を幾つかまとめるというものである。この解消方法を中山の研究では用いている。しかし、この解消方法では、まとめた動詞が複数の意味をもった語（多義語）になる可能性がある。その語の多義性を解消することは、まとめる前の動詞に戻すことであるためにできない。

多義語の問題においては、従来の研究では考慮されていない。単語の語義曖昧性解消の方法は、辞書で多義語を複数の意味にわけ、それら意味の中から適切なものを選ぶことを行うが、十分な精度がないため有効な方法でない。

本研究では、従来のスパース性解消の方法や、多義性解消の方法に問題があるために、主成分分析とよばれる統計的手法を用いて、スパースネスの問題と、配置先推定の研究で扱われなかった多義語の問題に対処する。また、主成分分析の性質を考慮した類似度を提案する。

本研究では、主成分分析の特性を通じて、従来の解消方法と同様な、動詞要素をまとめたり、動詞要素をわけたり、という操作が可能のため、スパースネスの解消と多義性の解消ができると思う。

実験では、EDR 概念体系辞書の「乗り物」に関する部分木を使用し、EDR 共起辞書から共起データを獲得した。実験方法は 10-fold cross validation を用いて、235 題の実験データを用意した。そして、それら実験データを多義語の問題とスパースの問題に分けて、シソーラスへ登録し評価した。実験で使う多義語の問題とは、乗り物の対象領域の中で、動詞要素に複数の意味がある場合とした。そうすると 213 題が多義語の問題となる。スパース

¹本研究では、名詞の多義性を考慮せず、動詞の多義性だけを考慮する。

ネスの問題とは、予め用意された正解ノードと実験データとのコサイン距離を計算し、コサインの値が0になるものとした。そうすると77題がスパースネスの問題となる。評価方法は、32種類の間ノードと未登録語との類似度計算の結果、第N位までに正解があるかどうかを調べる (N=1, 5, 10)。

実験の結果、主成分分析によって、スパースネスの問題と多義語の問題に効果が得ることができた類似度は、提案方法であり、主成分分析前と比べると、それぞれの問題に対して、最大で32.5%と9.4%の改善がみられた。

正解順位を10位まで認めると、スパースネスの問題に最も貢献する類似度の精度は、ユークリッド距離の41.6%であり、次に提案方法の32.5%であった。また、多義語の問題に最も貢献する類似度の精度は、提案方法の73.7%であり、次に、徳永の66.2%であった。

従来の研究では、未登録語の配置先を幾つか提示する方法をとる。本研究もこの方法に従うならば、従来の研究で提案された類似度と比べると、約10%の改善があった。