類似名詞のクラスタリングに基づく照応解析手法の提案

関口 友樹† 木村昌臣‡
芝浦工業大学大学院工学研究科† 芝浦工業大学工学部情報工学科‡

1.はじめに

近年,自然言語で記述された文に対して形態素解析や構文解析などの手法を用いた情報抽出や機械翻訳などの研究が盛んに行われている.しかし,これらの研究で扱う自然言語文では主語や目的語などが省略されることが多々あるため,そのような文章をそのまま解析に用いると結果の精度に影響が出る恐れがある.そこで,省略の補完を行うことを目的として,照応詞の検出や先行詞の同定などを行う照応解析が重要性を増している.本研究では照応解析の 1 手法として,関連した語をクラスタリングでまとめたシソーラスを構築し,動詞とクラスタ内の語の対応を見ることで照応解析の手掛かりとすることを考え,その前段階として本稿では単語間の関係からシソーラスを構築する実験について紹介する.

2.類似研究

辞書などのコーパスからシソーラスを構築する研究は過去にも多数存在する.中山らの研究[1]では wiki を利用して構築された百科事典である wikipedia を利用して特定の名詞と関連の高い語を見つけ出し,シソーラスを構築するシステムを構築している.この研究では wikipedia の語義文中の wikipedia 内へのリンクの数や距離などの構造を用いて関連性の計算を行うことでシソーラスを構築している.また鶴丸らの研究[2]では,国語辞典の見出し語と語義文の構造を利用して,単語間の上位/下位概念や同義語などを判別し,シソーラスを構築している.本研究では百科事典などの辞書から関連のある語をエッジで繋いだ複雑ネットワークを構築し,そのネットワークにクラスタ分析手法の 1 つであるスピングラスを用いたクラスタ分割手法を適用して,関連した語をまとめ,それをもとにシソーラスを構築する.その後,動詞と名詞クラスタの対応を見ることで照応解析の手掛かりとすることを考える.ここで,スピングラス分割手法とはクラスタ分析手法の 1 つであり,ネットワーク中で疎に分布しているエッジを見つけ出し,そこでネットワークを切り分けることでクラスタを構築する手法である.

3.提案手法

まず辞書から見出し語とその語義文を抽出する.そして,語義文に対し形態素解析を行い,語義文中の名詞を抽出する.ここで,見出し語と語義分中に存在する名詞は関連がある語同士であると考え,見出し語と語義文中の名詞との間とをエッジで繋ぐ. ただし,語義文中の複合名詞は形態素解析を行うと別々の名詞に分割されてしまうことがあるため,名詞が連続して出現している場合,それらを複合名詞として一つの名詞にまとめる処理を追加する.また一部の名詞は動詞「する」などを伴って動詞の働きをすることがある.よって,語義文中の名詞がサ変接続する場合は除外する. これらの処理をすべての見出し語と語義分中の名詞に対して行い,ネットワークを構築する.その後,ネットワークに対してスピングラス分割手法を用いることによってクラスタリングを行い,関連性の高い語をまとめる.そして,クラスタ内でハブとなっているノードは多くのノードに接続していることから抽象度の高い語であると考え,木構造の上位に配置することでシソーラスを構築する.

4.実験

実験はデータの取得が容易である点や網羅性が高い点などを考慮して公開されている百科事典データ[3]をコーパスとして用いた.この百科事典を用いて前節の手法でクラスタを構築した.語義文中の名詞の抽出には南瓜を使用し,クラスタの構築には統計解析ソフトの R を使用した.ただし,語義文に出現する回数が少ない語はクラスタの構築に影響をほとんど与えない,多くの語と接続しているハブの発見が容易になるなどの点から, 今回の実験では辞書中のすべての見出し語と語義文中の名詞で,語義文中に一定以上の頻度で出現する名詞のみを扱うこととする.

A proposal of anaphora analytical technique based on clustering the similar noun
† Tomoki Sekiguchi
† Graduate school of Shibaura Institute Technology
‡ Masaomi Kimura
‡ Shibaura Institute Technology

5.結果と考察

今回の実験では語義文中の出現頻度が 25 回以上の名詞を対象としたところ 847 個の名詞が抽出され,その名詞の組み合わせの数は 2172 組であった.これらのデータをクラスタリングしたところ 19 個のクラスタが構築された.表 1 は得られたクラスタの一部の例であり,図 1 は表 1 のクラスタ 11 をネットワークの視覚化を行うソフトである Pajek を用いて視覚化した図である.

結果を確認したところクラスタ 11 は「アメリカ合衆国」や「フランス」など多くが国名に関する語であり,同様にクラスタ 12 には主に宗教関係の語が含まれているなど,同じような意味を持つ名詞が同じクラスタに分割されやすいことが確認できた.これは,他のクラスタについても同様の結果が得られている.また,クラスタ 11 内部のノードとエッジのつながり具合を確認したところ,「国」や「共和国」,「首都」や「州」など国に関する上位/下位概念である語がハブとなっていた.これらの語は図 1 の中心近くに配置されているノードであり,「アメリカ合衆国」などのより具体的な名詞はこれらの語の周りに配置されている.このことから,今後クラスタ内のエッジのつながりを解析することで語の上位/下位概念などの語間の関係を知ることが出来ると考えられる. しかし,クラスタ 6 には「東京」や「江戸城」など日本の都市である東京に関する語と,「メートル」や「キログラム」 など単位に関する語があり,関連性の低いと思われる語が同じクラスタに分割されてしまった.このクラスタ 6 のエッジのつながりと百科事典の語義文を確認したところ,「江戸時代」のお金の単位である「両」や,都市の距離の単位などを仲介してエッジがつながったことが原因であった.このクラスタ内で再度クラスタリングを行ったところ,さらに 9 つのクラスタに分割された.表 2 は得られたクラスタの一部である.表 2 のようにクラスタ内で再クラスタリングをすることで,関連性の高いと思われる語を分割することが出来た.

表 1 構築されたクラスタ例

クラスタ番号	クラスタ内の名詞数	名詞例
クラスタ6	90個	単位,メートル,グラム,距離,単位記号 江戸城,江戸幕府,日本橋,東京...
クラスタ11	98個	アメリカ合衆国,イギリス,国,イラク... 通貨,通貨単位,憲法,国連,公用語...
クラスタ12	96個	イスラム教,インド,キリスト教,エルサレム チベット,チベット仏教,唐,仏,茶

表 2 クラスタ 6 内再クラスタリング結果

クラスタ番号	クラスタ内の名詞数	名詞例
クラスタ6-1	10個	江戸城,城,江戸,江戸幕府,徳川家康...
クラスタ6-2	17個	単位,ミリメートル,グラム,センチメートル...
クラスタ6-3	10個	日本橋,橋,東海道,鉄道,道路

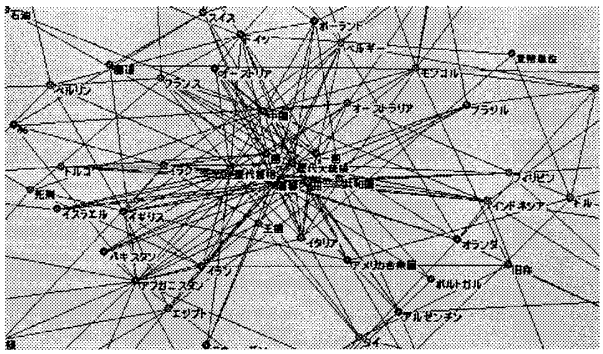

図 1 クラスタ 11 のネットワーク構造

今回の実験では「花」や「動物」といったような抽象的な語がハブにならないという傾向が見られた.本来このような語は多くの語の説明として使われることが多く,実際に「花」という名詞は 170 個の見出し語の語義文中で使われているが,今回は語義文の出現頻度が 25 回未満の名詞を排除したため,「花」や「動物」を語義文に持つ名詞が排除されたことが原因だと思われる.しかし,このような語もクラスタリングする必要があるため,今後は出現回数以外で語を絞る処理が必要であると考えられる.

6.まとめ

本稿では関連した語をまとめたシソーラスを照応解析に適用させる前段階として,スピングラス分割手法を用いた辞書からのクラスタの構築実験を行った.今回の実験では見出し語と語義文中のすべての名詞を関連する語としてネットワークを構築したが,利用する辞書によっては語義文中に見出し語と関係性の低い語が現れる場合もあるため,これに関する考察が必要である.また,今回の実験はネットワークを無向グラフとして作成したが,これを有向グラフにすることで,単語間の関係の同定などに利用することが出来ると考えられる.また構築したシソーラスと動詞との対応の取り方に関する考察も今後の課題である.

参考文献

[1] 中山浩太郎,原隆浩,西尾章治朗:
大規模 Web 事典からのシソーラス辞書構築,日本データベース学会 Letters, Vol. 5, No. 4, pp. 41-44 (2007)

[2] 鶴丸弘昭,竹下克典,伊丹克企,柳川俊英,吉田将:
国語辞典情報を用いたシソーラスの構築について,情報処理学会研究報告,1991-NL-83(1991)

[3] 私立 PDD 図書館/百科事典
http://www.cnet-ta.ne.jp/p/pddlib/japanese/index.htm