1C3-1

マイクロブログを利用した動的なシソーラス自動構築手法の提案

Proposal of a Method for Automatic Construction of dynamic thesaurus by MicroBlog

辻井　由佳　　西山　裕之
Yuka Tsujii　　Hiroyuki Nishiyama

東京理科大学理工学研究科経営工学専攻
Graduate School of Science and Technology, Tokyo University of Science

Query expanded by thesaurus is the one of efficient information search technology on Web. In these days, anyone can publish information based on penetration of the internet, many new words are born. Therefore thesaurus needs to be updated, however manual updates causes many troubles. This study proposes a method for automatic construction of thesaurus used MicroBlog as information source. Concretely, the method uses Co-Occurrence rate. I believe that using MicroBlog can establish the method to construct thesaurus of new words and frequently-used words.

1. はじめに

近年，ソーシャルネットワーク (SNS) やブログの台頭によって一般のユーザも情報発信をすることが容易になり，情報を得るためにインターネットを利用するのが一般的になった [総務省 11]. その一方で情報の氾濫はインターネット上の情報取得を難しくしており，必要な情報を得るためにはいかに適切な検索を行うかが問題となる．そのようなウェブ上の情報検索を効率的にする方法としてシソーラスを使用したクエリの拡張がある．シソーラスとは同義語から広義・狭義の類義語をまとめた辞書を指す．これを用いてクエリ拡張を行えば，日本と検索した際に Japan といった表現も同時に検索できるため，より多くの情報が取得できる．また，シソーラスは検索以外の言語処理にも利用することが可能である．

このようなシソーラス構築の研究としてユーザの検索履歴を使用した安川ら [安川ら 04] のものがあるが，ユーザ毎のシソーラス構築が行える一方で新規に検索する語については対応していない．渡部ら [渡部ら 11] の研究ではある関係にある語ペアの正例となるデータが必要とされることがわかっている．

また，近年では誰でも情報発信が可能になったことで，新たな言葉が多数生まれていたり，表記が統一されていないことによる表記ゆれ問題があり，シソーラスは更新されていくことが必要とされる．しかし，常に変化していくシソーラスの構築は手動で行うには手間がかかりすぎるという問題が存在する．そこで，本研究では自身の状況などを短文で書き込むためのツールで，リアルタイム性の高いマイクロブログを情報ソースとしてシソーラスの自動構築を行う手法を提案することを目的とする．具体的には二つのキーワードが同一文書中に現れる確率を示す語の共起度を使用し，シソーラスを手に入れたい言葉についてマイクロブログ上から同義語や類義語を取得する．このようにマイクロブログを使用することで，新たに生まれた語やより使われる語のシソーラス構築が行えると考えられる．

2. 提案手法

提案手法ではシソーラスを得たい単語についてマイクロブログで検索し，それに基づいてシソーラスを構築する．今回，マイクロブログとしてもっともよく知られている Twitter[*1] を使用した．

まず今回の提案手法の手順を示す．

1. シソーラスを得たい語（中心語）でマイクロブログを検索する
2. 検索結果をクラスタリング
3. 中心語と共起度の高い単語 (共有語) を探す
4. 共有語で検索クエリを作成
5. 共有語クエリで検索
6. 手順 5 の検索結果から共有語と共起度が高い語を中心語のシソーラスと定義する

シソーラスを得たい語を中心語と呼び，また，中心語と共起度の高い語を共有語と呼ぶことにした．ここで，クラスタリングには前年度 [辻井ら 12] の研究を使用した．クラスタリングによって共通の話題にまとめることができ，より共通する語が見つけやすくなり、共有語候補を減らすことができると考えている．

次に各手順について詳しく説明していく．

2.1 共起度

共起度とはある単語とその単語が同じ文書に現れる確率を示す．具体的に，語 A と語 B の共起度は以下の式 1 で示すように定義される．

$$\text{共起度} = \frac{\text{語 } A \text{ と語 } B \text{ が同時に使われる文書数}}{\text{語 } A \text{ もしくは語 } B \text{ が出現した文書数}} \quad (1)$$

本論文におけるシソーラス構築を簡略化すると以下の図 1 のようになる．

たとえば文書中にある語 A と共起度の高い語 B と語 C があり，他の文書である語 D と共起度の高い語 B と C が存在した場合は語 A と語 D は似た意味を示す同義語，もしくは類義語だと考えられる．この考え方に基づき，まず中心語と共起

連絡先: 辻井由佳，東京理科大学理工学部，千葉県野田市山崎 2641，j7412616@ed.tus.ac.jp

*1 Twitter：https://twitter.com/

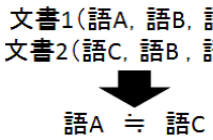

図 1: シソーラス構築の簡略化イメージ

度の高い共有語を見つけ出す．これに基づいて共有語の検索クエリを構築し，その検索クエリでマイクロブログ内から追加で情報を取得する．すでに取得してあるものを削除し，残ったツイートから，共有語と共起度の高い語を探す．この共起度が高い語が中心語の同義語もしくは類義語であると考えられる．

2.2 カテゴリ

本研究では，より多くの情報を集めるために語のカテゴリを使用した．語のカテゴリの取得にははてなキーワード[*2]を使用した．はてなキーワード API の使用により，登録されている単語であれば web,book,science などの 20 種のカテゴリを得ることができる．

本論文では同一ツイート中の共有語で同カテゴリの語は一つにまとめることにした．短いツイートの中で使われる同一カテゴリには強い関連性があると考えられ，また，そのように共有語の条件を広げることでより多くの追加情報が取得できると考えている．

3. 設計

今回，Twitter からのデータ収集には Twitter 非公式ライブラリである Twitter4J[*3]を使用し，文を単語単位に分解する形態素解析にはキーフレーズ抽出[*4]を利用し，これを用いて文中の単語を取得する．これにより，最近作られた単語など辞書に登録されていない用語についても特徴として利用することができる．

本システムのシステムフローを図 2 に示し，各処理について確認する．

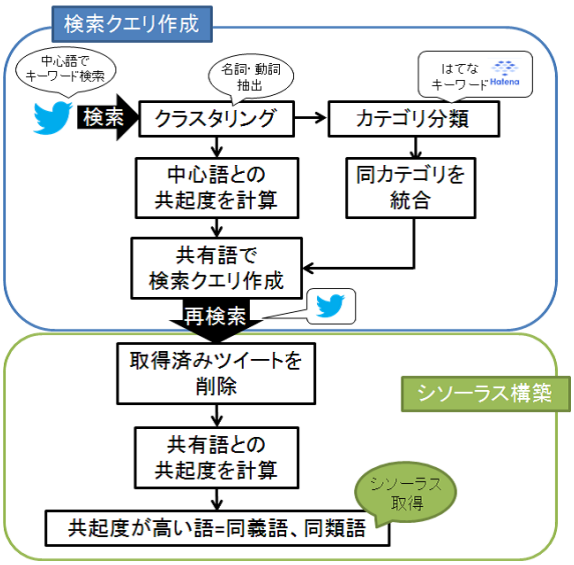

図 2: システムフロー

*2 はてなキーワード：http://d.hatena.ne.jp/keyword/
*3 Twitter4J：http://twitter4j.org/ja/index.html
*4 Yahoo!キーフレーズ抽出：http://developer.yahoo.co.jp/

3.1 共有語の検索クエリ作成

まず，共有語による検索クエリの作成について説明する．中心語をマイクロブログのキーワード検索し，その結果でクラスタリングする．クラスタリング時に得られた単語のうち，名詞と動詞を共有語候補とした．共有語候補をカテゴリに分け，共有語候補と中心語の共起度を式 1 に基づいて計算する．共起度の高い単語と，その単語があるカテゴリに属している場合はそのカテゴリに属する語も利用して検索クエリを作成する．また，このとき，なるべく関連した話題を得るために共起度が高い順に 2 語以上を選択することにした．

以下に本システムでの検索クエリ例を示す．

(半島|韓国|ソウル) & 台風 & 接近

図 3: 検索クエリ例

ここで，”| ”は ”または ”，”& ”は ”かつ ”を意味する記号である．上記の例で検索をかけると ”半島、韓国、ソウルのうちいずれかの語、かつ台風と接近 ”が文中に存在するツイートを取得できる．

3.2 シソーラス構築

検索クエリ作成した後，中心語の同義語・類義語となる言葉を探す．

まず，検索クエリでマイクロブログからデータを取得する．このとき，すでに取得したものに関しては削除し，新たに得られたデータのみを対象とした．そのデータ内の単語のうち，名詞のみを取得して共起度を計算する．その中から共有語と共起度が高い単語を中心語の同義語・類義語とみなす．

4. おわりに

本研究ではマイクロブログを情報ソースにシソーラスの構築を目指す．マイクロブログの使用によって，より使われる語，新しくできた言葉のシソーラスが取得できると考えており，そこから共起度によってシソーラスを求める手法を提案した．クラスタリングによってある程度関係する話題を絞り込むことで，シソーラスとなる同義語・類義語候補を減らし，中心語と共起度の高い共有語を求めてシソーラスを構築するというのが本研究の提案手法である．

今後は語のカテゴリを使った影響について考察し，より新しい語についてシソーラスが取得できているか，またシソーラスとしてふさわしいかどうかを評価していきたいと考えている．

参考文献

[総務省 11] 総務省:情報通信白書 2011 年度版，2011

[安川ら 04] 安川美智子，山田篤:Web 検索エンジンを用いた用語検索履歴からのシソーラス自動構築，日本データベース学会 letters 3(1)，pp.105-108，2004

[渡部ら 11] 渡部啓吾，Danushka Bollegala，松尾豊，石塚満:検索エンジンを用いた関連語の自動抽出，日本知能情報ファジィ学会誌 23(5)，pp.739-748，2011

[辻井ら 12] 辻井由佳，西山裕之:マイクロブログの特徴を考慮した文書クラスタリング手法の提案と実装，人工知能学会全国大会 (第 26 回) 論文集，4I1-R-9-2，2012