

質問応答 Web サイトからの関連語ネットワークの自動抽出

吉田 知訓[†] 間瀬 心博[†] 北村 泰彦[†]

[†] 関西学院大学 理工学部 情報科学科

あらまし 本論文では、Web 上の質問応答サイトの口コミ情報から、単語を抽出し関連するキーワードを結んだ関連語ネットワークを構築する手法を提案する。質問応答サイトの回答文には質問文に関連する事柄が記述されていることを用い、質問文と回答文に含まれる同じ属性の単語を関連語として抽出し、関連度の高い語を結んで関連語ネットワークを構築する。関連度として共起頻度、Jaccard 係数、Simpson 係数を用いて関連度ネットワークを構築し評価を行った。

キーワード 関連語ネットワーク、シソーラス、共起情報、Web マイニング

Automatic Extraction of Related Term Networks from FAQ Sites

Tomonori YOSHIDA[†], Motohiro MASE[†], and Yasuhiko KITAMURA[†]

[†] Department of Informatics, School of Science and Technology, Kwansei Gakuin University

Abstract This paper describes a method for building a network of related terms which shows relationships of the terms from word-of-mouth information of FAQ sites on the Web. We extract terms which have a same attribute as related terms and build the network of related terms by connecting terms of high relevance. We build networks of related terms by calculating relevance of terms using co-occurrence frequency, Jaccard and Simpson measure and evaluate the extracted networks.

Key words

1. はじめに

Web2.0 の時代といわれるようになり、人々の Web の利用の仕方が変わり、人々の情報との向き合い方が変化している。インターネットが普及し始めた初期は、情報発信力の強いマスメディアや大企業、有名人による一方通行なメディアであったが、近年は Blog や Wiki に代表されるように、これまで情報の受け手であるユーザが情報を発信する立場になり、双方向のメディアとなろうとしている。それに伴い、Web 上での口コミ情報の影響力も大きくなっている。Web 上の口コミ情報は、情報の伝播が早く、発信者が多いという特徴がある。口コミ情報を参考にして商品を購入する消費者が増え、また企業が口コミ情報を利用してマーケティングを行うようになり^(注1)、Web 上の口コミ情報の注目度は上がっている。

口コミ情報を取り扱っている Web サイトの一つに「教えて!goo」^(注2)や「Yahoo!知恵袋」^(注3)といった質問応答 Web サイトがある。これらは「食べ歩き情報」や「携帯電話」等といっ

たカテゴリ別に分かれており、ユーザー同士が質問や回答をすることによって情報の蓄積を行っており、大量の質問応答文がある。これらから情報を抽出し、体系化することによって有用な情報源として再利用できる [12]。一般的に回答文には質問文に関連した事柄が記述される。この性質を利用し、質問文に書かれている事柄と回答文に書かれている事柄を関連付ければ、関連語として抽出できる。また、質問文には広義的な意味の単語が見られ、回答文には狭義的な意味の単語が見られるため、上位下位関係にある関連語が抽出することもできる。さらに関連する単語同士を結び可視化することで関連語ネットワーク [9] を構築する。

Web 上の口コミ情報を利用しているため、その時々流行や関心のある事柄を反映した関連語ネットワークが構築でき、中心となるものや、時期ごとの移り変わりをみることにより、関心の高い事柄や流行がわかる。また、関連語のつながりを可視化することによって、発想支援やチャンス発見の手がかりになることが期待される [8]。

本研究では、質問応答 Web サイトを情報源として用いて関連語を抽出し、関連語ネットワークを構築する手法を提案する。提案手法は、質問回答文の特徴を利用して、特に上位下位関係にある関連語の抽出を行い、関連語から関連語ネットワークを

(注1): インターネットによるクチコミ現象の動向, <http://www.net.c.dendai.ac.jp/~maru/>

(注2): <http://oshiete.goo.ne.jp/>

(注3): <http://chiebukuro.yahoo.co.jp/>

構築する。第2章では、関連語抽出の関連研究について述べる。第3章では、本研究で情報源として使用する質問応答 Web サイトとその性質について述べる。第4章では質問応答 Web サイトから関連語を抽出する手法について述べる。第5章では共起頻度、Jaccard 係数、Simpson 係数を関連度として用いて関連語ネットワークを構築し、各関連語ネットワークと抽出された上位下位関係にある関連語の割合について評価を行う。第6章で本研究のまとめを行う。

2. 関連研究

これまで、文書中から語の意味的な関係や関連語を抽出する研究は多く行われている。それらの多くの研究は大量のコーパスから構文パターンのマッチングによって単語間の関係を抽出を行っている [2], [3], [6]。Hearst は大量のコーパスから「such as」の構文パターンのマッチングによって単語間の上位下位関係を抽出している [3]。安藤らは新聞記事を対象にして、「トマトなどの野菜」のような連体修飾表現を用いて、名詞の下位概念を自動的に抽出を行っている [6]。これらの手法の問題点としては、対象のコーパスに構文パターンが現れる頻度が低いため、広範囲の単語間の関係を抽出することが難しいという問題点がある。

そのため、構文パターンによる抽出を行わずに、コーパス内での単語の出現頻度に注目して、関連語の抽出を行う研究が行われている [9], [10]。榊らは Web 上での語の共起情報を利用したカイ二乗値による語の関連度を求め、関連度による語のネットワークを作成しクラスタリングを行うことで、シソーラスの自動的な構築を行っている [9]。廣川らは、文書群から特徴語を抽出し、さらに特徴語の出現頻度を解析することで単語間の上位下位概念の抽出を行い、関連語間の関係を概念グラフとして可視化している [10]。大澤らは、文書中の単語の出現頻度と共起関係に注目することでグラフを作成し、さらにそのグラフから文書の主張点を把握し、単語や単語間の関係を抽出する手法である KeyGraph [8] を提案している。

また、使用する情報源の特徴を利用する研究も行われている [1], [4], [5]。荒牧らは辞書の特徴を用い、医療分野においての単語間の関連を示した [5]。辞書の定義文は見出し語を説明したものであり、日本語においては末尾に見出し語の上位語や同意語が書かれている場合が多いという辞書特有の性質を用いている。Chen らは Web サイトの特徴に注目して、特定ドメインの Web サイト、関連する Web ページのリンクからリンクグラフを作成し、Web ページにリンクのアンカーテキストによるアノテーションを行うことで、特定ドメインにおけるシソーラスを自動構築を行っている [1]。新里らは、大量のコーパスから獲得した名詞と動詞の係り受け関係に加えて、Web ページの HTML タグを利用した HTML 文書の構造等を利用することで、HTML 文書から単語間の上位下位関係を自動抽出を行っている [4]。

これらの研究に対して、本研究では情報源として質問応答 Web サイトを用いる。質問応答 Web サイトでは、回答文には質問文の記述に対応する内容やより狭義的な内容が記述され

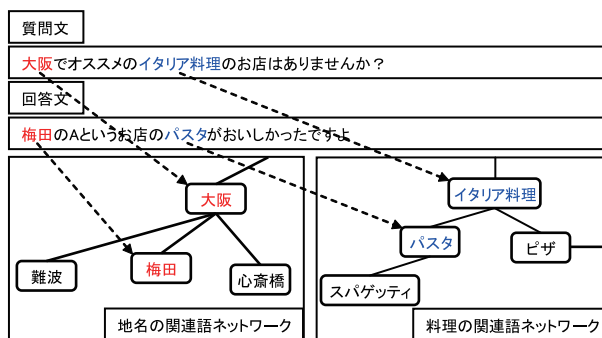


図1 質問応答文からの関連語抽出

Fig. 1 Extraction of related terms from question answering sentence

ているという特徴を利用している。質問応答文から用語データベースと分類ルールを用いて単語とその属性を抽出し、同属性の単語を関連付けることで、上位下位関係に注目して関連語の抽出を行う。また、分類ルールを用いているため、用語データベースにない未知語に対応することが可能である。

3. 質問応答 Web サイト

本研究では情報源として質問応答 Web サイトを利用し、関連語ネットワークを構築する。質問応答 Web サイトは、情報交換に特化した Web サイトであり、利用者間で質問・回答を行うことで情報の蓄積を行う Web サイトである。「教えて!goo」は、最もよく利用されている質問応答 Web サイトである^(注4)。「教えて!goo」は、ライフ・エンターテインメント・旅行・レジャー・生活などの 11 のカテゴリに分別されており、そのカテゴリもさらに詳細なサブカテゴリに分割されている。ユーザは自分が知りたい事柄に合ったカテゴリで質問を行い、寄せられた回答にポイントを与えることができる。このポイントは集計され、獲得ポイントの多い回答者は表彰され、抽選で商品を得ることができる。このような仕組みのため、Web 上に多くある一般的な掲示板に比べ、良質な回答が得ることができる。

本研究では関連語を抽出するために、質問応答 Web サイトの特徴を利用する。質問応答 Web サイトでは、質問文と回答文に分かれており、また一般的に質問文には広義的な意味の単語がみられ、回答文には狭義的な意味の単語が見られる。質問文に出現する単語と回答文に出現する単語が同じ属性の場合、それらの単語を関連付けることによって単語間の上位下位関係を抽出できる。例えば、図1のような「大阪でオススメのイタリア料理のお店はありませんか?」という質問文と、「梅田のAというお店の Pasta がおいしかったですよ」という回答文の場合には、地名に関しては大阪と梅田、料理に関してはイタリア料理と Pasta という情報が含まれている。この質問応答文から単語とその属性を抽出し、同じ属性の単語を関連付けることで、〈大阪, 地名-梅田, 地名〉, 〈イタリア料理, 料理-Pasta, 料理〉という関係を抽出できる。また、大阪は梅田の上位

(注4): 「質問サイトの利用状況」に関する調査結果, <http://research.goo.ne.jp/database/data/000262/index.html>

質問	
QNo.2571273	関西で美味しい蕎麦が食べれるところを教えてください。
質問者: hixlxix	蕎麦にはまっています。新蕎麦の季節。ぜひとも美味しい、せいろ
回答	
ANo.6	和歌山市の畑屋敷という場所にある「信濃庵」という蕎麦屋は美味しいです。とこぎる蕎麦は本当に美味しいです。
回答者:	

図 2 質問応答の例 [教えて!goo]

Fig. 2 Example of question answering

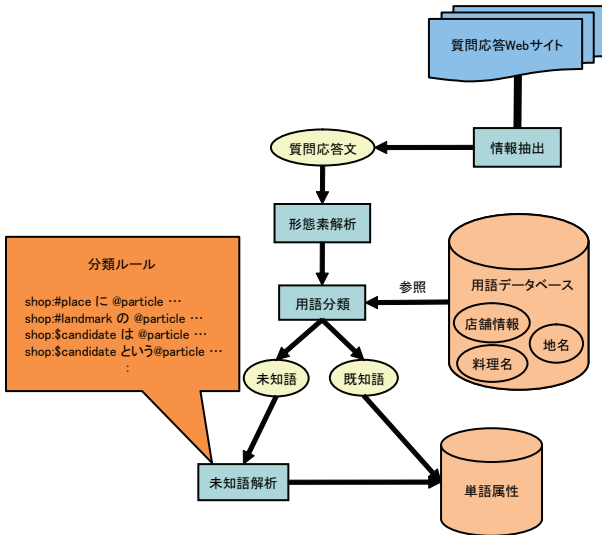


図 3 単語と属性の抽出

Fig. 3 Extraction of terms and attributes

語, イタリア料理はパスタの上位語とすることができる。

4. 質問応答 Web サイトからの関連語ネットワークの構築

4.1 単語と属性の抽出

関連語を抽出するためには, 単語とその属性が必要になる。そのために, 質問応答文から単語とその属性を抽出する。質問応答 Web サイトである「教えて!goo」でのレストラン情報に関する質問応答のやり取りを, 図 2 に示す。

この例には「蕎麦」という料理名, 「和歌山」「畑屋敷」という地名が含まれており, <蕎麦, 料理名>, <和歌山, 地名>, <畑屋敷, 地名> という単語とその属性を抽出することができる。本研究における単語とその属性の抽出手順を図 3 に示す。

まず, 質問応答 Web サイトから質問文と回答文を抽出し, 形態素解析を行い品詞情報を付与する。質問応答 Web サイトにおいて, 「「辣子鶏」という料理」というようにキーワードとなる語句は「」でくくられることが多く, これらを候補語とする。次に形態素解析の結果と用語データベース(表 1)を照合し, 用語データベースと一致した名詞や候補語に, 地名や料理名といった属性を付与する。用語データベースは, 分類語彙表や, 郵便局のホームページから地名情報を抽出するなど Web 上の情報から作成した。しかし, 用語データベースに含まれていない単語もあるため, それらを未知語とし分類ルールを用いて分類する。

表 1 用語データベース

Table 1 Terms database

ジャンル	料理名	駅名	ランドマーク	地域
語数	7,795	9,146	11,518	132,374
用語例	からあげ そば :	新宿 浅草 :	お台場 武道館 :	池袋 六本木 :

質問	
QNo.17239	ルミナリエに行くんですが...
質問者: none-blue	17日に神戸のルミナリエに行く予定ですが, 県外なので美味しいお店がわかりません。困り度: あんまり高くなくて, 美味しくて, 出来れば雰囲気いいお店を教えてください。
回答	
ANo.1	三宮にある『エスカレー』というフレンチのお店は美味しいです。特にランチが安くオススです。ランチのコースは1300円からです。夜は行ったことがありませんが, フルコースは2000円からのようです。
回答者: kob	

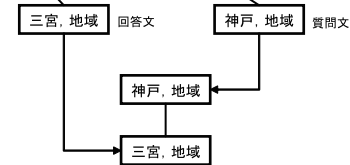


図 4 単語の関連付け

Fig. 4 Term association

例えば, あまり人に知られていない料理名を挙げるとき, 「という料理」という言い回しがされる。ここから, 形態素解析の結果が「X(未知語) + という(助詞) + 料理(名詞)」という並びの場合, その未知語 X は料理名と判断することができる。これを分類ルールで表すと以下のように書くことができる。

food: \$unknown という@particle 料理@noun

unknown は未知語, particle は助詞, noun は名詞を表し, \$ は抽出対象の単語, @ は品詞, # は集合を表し, このルールに一致するものがあれば未知語を料理名と判別する。このルールを適応すると, 「四川料理屋でよく食べていた「辣子鶏」という料理があります」という一文から, 「辣子鶏」という料理名が得られる。

上記のような分類ルールを複数組み合わせることで, 単語とその属性を抽出する。本研究で使用する分類ルールを以下に示す。

- (1) food: \$unknwon, という@particle, 料理@noun
例) 四川料理屋でよく食べていた「辣子鶏」という料理
- (2) food: \$unknwon, particle, verb
例) 東京でシュラスコを食べたい
- (3) food: NUM, 円@noun, *, \$unknown
例) 1800 円のふかひれの姿煮
- (4) food: #shop, の@prticle, *, \$unknown
例) マクドナルドのビッグマック

4.2 関連語の抽出

上記の手法で抽出した単語とその属性をもとに関連語を抽出する。図 4 の場合には, 質問文から <神戸, 地域>, 回答文から

ら<三宮, 地域>という情報が抽出され, <神戸-三宮>と関連付けられる. 一つの質問文には複数の回答文がつくことがあるため, 回答文の数だけ質問文と回答文の組を作る. 質問文と回答文に同じ属性の単語があれば関連付けをする. 全ての質問応答文から同様に関連付けを行い, 関連付けられた単語の組のうち関連度の高い組を関連語として用いる.

4.3 関連度の計算

本研究では, 単語間の関連性の指標として用いられることの多い共起頻度, Jaccard 係数, Simpson 係数を関連度として用いる [7], [9]. 単語 A を含む関連語が出現した数を $|A|$, 単語 B を含む関連語が出現した数を $|B|$, 単語 A と単語 B が同時に出現した数を $|A \cap B|$, 単語 A または単語 B が関連語に出現した数を $|A \cup B|$ とすると, それぞれの関連度は次式で求めることができる.

- 共起頻度

$$|A \cap B| \quad (1)$$

- Jaccard 係数

$$\frac{|A \cap B|}{|A \cup B|} \quad (2)$$

- Simpson 係数

$$\frac{|A \cap B|}{\min(|A|, |B|)} \quad (3)$$

Simpson 係数では, $|A|$ と $|B|$ に大きな差があるとき, 正しく関連度を測れない場合がある. $|A|$ が 1000, $|B|$ が 1, $|A \cap B|$ が 1 とすると, $\min(|A|, |B|)$ が 1 となり, 同時に現れた数が 1 にもかかわらず値は大きくなる. 出現数の低い語はノイズの可能性が高いため, 値が大きくなることは望ましくない. この問題を解決するため, $|A|$ と $|B|$ の値が閾値以上のものであった場合に Simpson 係数を求め, 閾値以下の場合には関連度を 0 とする手法がある [11]. 本研究では閾値を 10 として Simpson 係数を計算する.

4.4 関連語ネットワークの構築手順

関連語ネットワークを構築する手順を以下に示す.

- (1) 質問応答 Web サイトから質問応答文を獲得する.
- (2) 質問応答文に出現する単語とその属性を抽出し, 同じ属性の単語の組を作成する.
- (3) 全ての単語の組の関連度を計算する.
- (4) 関連度の高い上位 50 件の単語の組を用いて, 各単語をノードとして, 組になっている単語間にエッジを張り, 関連語ネットワークとして構築する.

5. 関連語ネットワークの構築

本研究では「教えて!goo」内の食べ歩きカテゴリの質問応答文約 8000 件に対して, 共起頻度, Jaccard 係数, Simpson 係数を関連度として用いて関連語を抽出し, 地名と料理名の関連語ネットワークを構築する. 構築した各関連語ネットワークに含まれる全ての関連語の上位下位関係にある関連語の割合を評価した.

5.1 共起頻度に基づく関連語ネットワーク

共起頻度を関連度として用いて, 関連語ネットワークを構築した. 地名の関連語ネットワークを図 5 に, 料理名の関連語ネットワークを図 6 に示す.

図 5 の地名の関連語ネットワークでは, 東京や大阪, 京都, 神戸といった都市を中心に下位の都市や近隣の都市のクラスタが確認できる. しかし, <東京-大阪> や, <東京-神戸> といった有名都市同士も関連語として抽出されているが, これらは望ましくない. 例えば, 「大阪から東京にいくのですがおすすめの店はありますか」といった質問文に対し「東京では A というお店がいいですよ」という回答がされた場合には, <東京-大阪> といった関係が抽出される. しかし, この場合には東京と大阪の関連性は低いため, 関連度が高くなるのは望ましくない. このように, 有名な都市は質問の本質とは無関係に出現する頻度が高く, 共起頻度をそのまま関連度とすることは不適切だといえる.

図 6 の料理名の関連語ネットワークでも地名と同様に, 頻出語であるラーメンやパンを中心にエッジがのびているが, <ラーメン-寿司> といった関連性の低い語が関連語として抽出されている.

5.2 Jaccard 係数に基づく関連語ネットワーク

Jaccard 係数を関連度として用いて, 関連語ネットワークを構築した. 地名の関連語ネットワークを図 7 に, 料理名の関連語ネットワークを図 8 に示す.

図 7 の地名の関連語ネットワークでは関連語の組が散在しており, 広い地域においての関連語が抽出されている. しかし, 東京が含まれていない, 大阪に関係した地名が梅田しかでてきていないなど, 人々の関心の高い地域が表せているとはいえない. これは, Jaccard 係数には単独で出現する頻度が高い語ほど値が低くなるという性質があるからである. 例えば, 「東京」が 1000 件, 「渋谷」が 50 件あり, 「東京」と「渋谷」同時にみられたのが 50 件だったとすると, 「渋谷」はすべて「東京」と同時に出現しているのに値としては小さなものとなる.

図 8 の料理名については, ハンバーガーやインド料理に関する関連語の小規模なクラスタが散見できた. しかし, 地名のときと同様にひとつのノードからのびるエッジの数は少ない. また, ラーメンやパンといった出現頻度の高い語を含む関連語についても抽出されていない.

5.3 Simpson 係数に基づく関連語ネットワーク

Simpson 係数を関連度として用いて, 関連語ネットワークを構築した. 地名の関連語ネットワークを図 9 に, 料理名の関連語ネットワークを図 10 に示す.

図 9 の地名の関連語ネットワークでは, 東京, 大阪, 京都といった頻出都市を中心にネットワークが構築されている. 共起頻度をそのまま関連度とした場合, <東京-大阪> といった, 地理的に離れている有名都市が関連語として抽出されたが, Simpson 係数を関連度とした場合は抽出されていない. これは, 有名都市は出現頻度が高いため共起頻度は大きくなるが, Simpson 係数を用いた場合には, 有名都市は単独で出現する数も多いため, 分母が大きくなって関連度も低くなるためである.

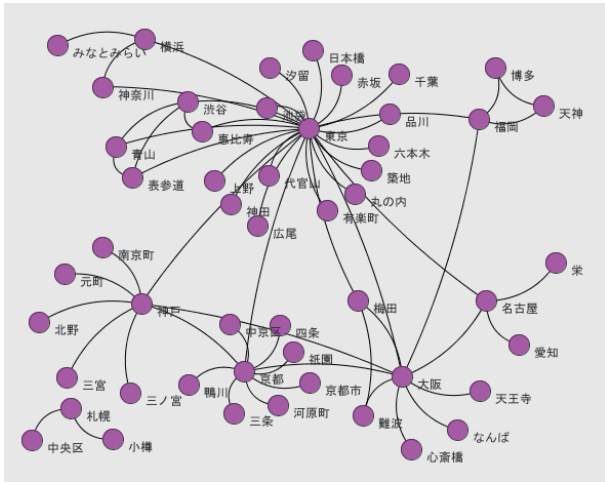


図 5 共起頻度を関連度とした地名の関連語ネットワーク
Fig. 5 Network of related terms with co-occurrence frequency measure(Place)

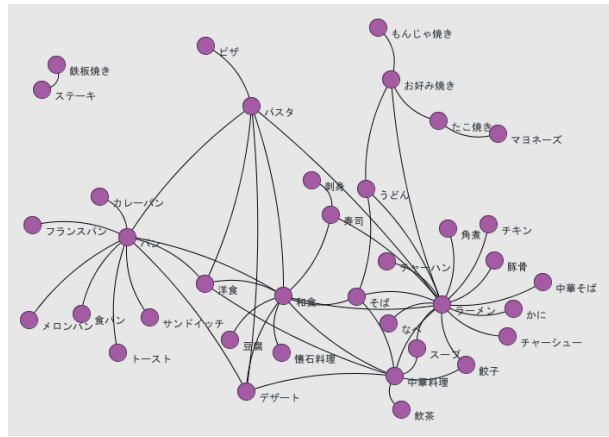


図 6 共起頻度を関連度とした料理名の関連語ネットワーク
Fig. 6 Network of related terms with co-occurrence frequency measure(Food)

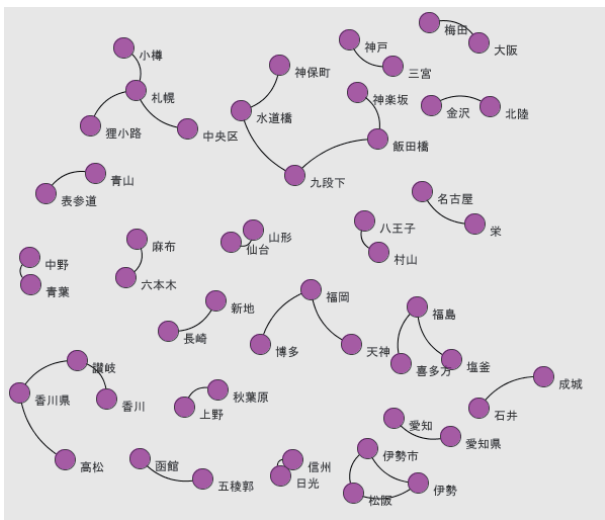


図 7 Jaccard 係数を関連度とした地名の関連語ネットワーク
Fig. 7 Network of related terms with Jaccard measure(Place)

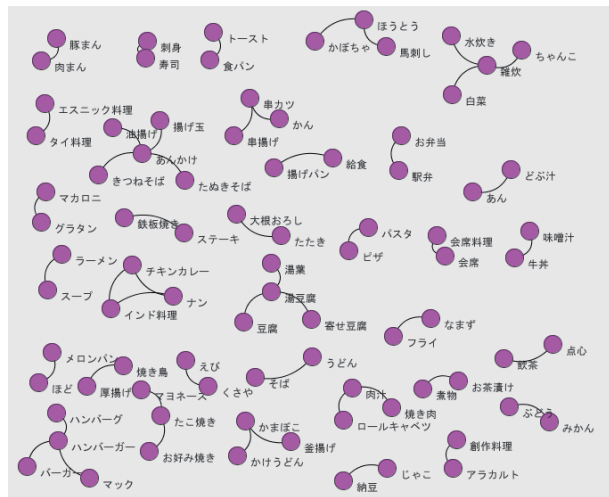


図 8 Jaccard 係数を関連度とした料理名の関連語ネットワーク
Fig. 8 Network of related terms with Jaccard measure(Food)

図 10 の料理名の関連語ネットワークでは、パンやラーメンといった頻出語を中心に、<パン-メロンパン>、<ラーメン-豚骨ラーメン>といった上位下位関係や<ラーメン-チャーシュー>という料理とその構成要素、<ラーメン-餃子>といった一緒に食べる機会の多い料理名が関連語として抽出された。ピザ、パスタ、スパゲッティといったイタリア料理についてのクラスタ、明石焼き、たこ焼き、お好み焼き、もんじゃ焼きといった粉物料理についてのクラスタが確認できた。

5.4 上位下位関係の関連語の割合

各関連度を用いて構築した関連語ネットワークにおいて、上位下位関係にある関連語が含まれる割合を表 2 に示す。表 2 より、地名、料理名の関連語ネットワークにおいて、Simpson 計数を関連語として用いた場合が、上位下位関係の関連語を多く抽出できていることがわかる。共起頻度による関連度を用いた場合には、単独で出現する頻度の高い語の値が大きくなりやすく、出現頻度の高い単語同士を関連語として抽出することにな

る。Jaccard 係数においては、単独で出現しやすい語を含む関連語ほど関連度の値は小さくなるため、東京や大阪、ラーメンやパンといった多く話題に上っている語を含んだ関連語の関連度が低くなる。出現頻度の高い語ほど、それに関する話題が語られやすく、それだけ上位語や下位語が出現しやすくなる。そのため、Jaccard 係数を用いた場合の上位下位関係にある関連度の割合が低くなるという原因となった。一方、Simpson 係数では出現頻度の低い語を基準に算出するので、出現頻度の高い語を含む語の関連度も高くなり、上位下位関係を含む割合が高くなった。地名に比べ、料理名の関連語ネットワークに含まれる上位下位関係が少なかったのは、質問文や回答文に「お好み焼きやたこ焼きが食べたい」といったように、似たジャンルの料理が同時に現れたり、「A というお店のうどんは掻き揚げや天ぷら、かまぼこをトッピングできます」というような料理名とその構成要素が出てくることが多かったためである。

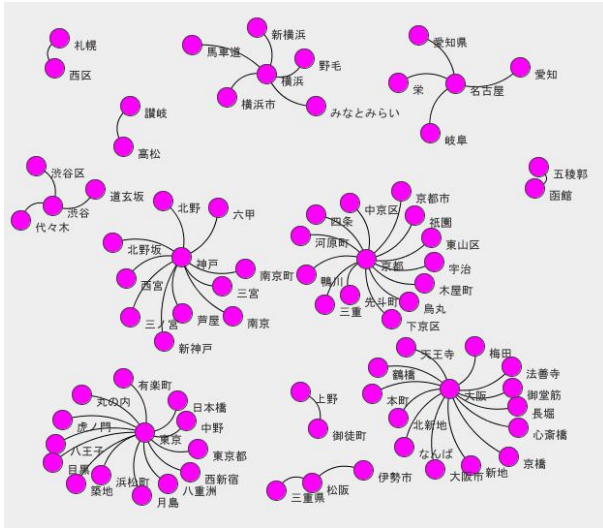


図 9 Simpson 計数を関連度とした地名の関連語ネットワーク
Fig. 9 Network of related terms with Simpson measure (Place)

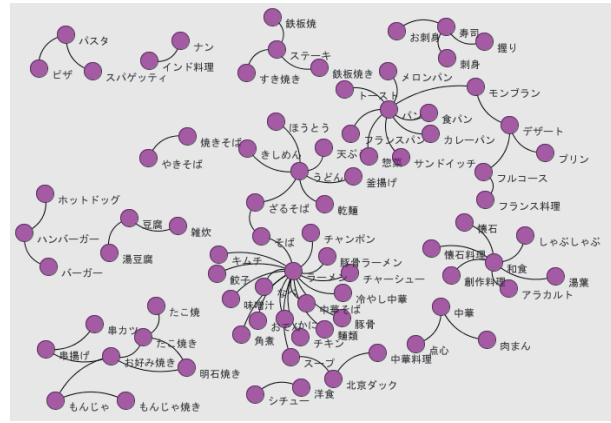


図 10 Simpson 係数を関連度とした料理名の関連語ネットワーク
Fig. 10 Network of related terms with Simpson measure (Food)

表 2 関連語ネットワークに上位下位関係の関連語が含まれる割合
Table 2 Rate of hierarchical terms in the network of related terms

	共起頻度	Jaccard 係数	Simpson 係数
地名	78 %	54 %	88 %
料理名	40 %	28 %	56 %

5.5 今後の課題

本研究では、質問文と回答文に含まれる同じ属性を持つ単語を関連語として抽出している。そのため、関連語ネットワークは同属性の単語のみで構成される。地名と料理名の両方を含んだ関連語ネットワークなど、複数の属性を単語によって構成される関連語ネットワークが構築できれば、より有用なものになると考えられる。地名と料理名が関連付けられれば、口コミ情報による各地の名産品マップを構築することもできる。また、関連語間にどのような関連があるのかを特定することが課題となる。

6. まとめ

本研究では、情報源として口コミ情報が多く含まれる質問回答 Web サイトを用いて、関連語を抽出し関連語ネットワークを構築する手法の提案を行った。提案手法は、質問回答文の特徴を利用して、特に上位下位関係に注目した関連語の抽出を行う。

「教えて goo!」の食べ歩きカテゴリの質問回答文を対象に、関連度として共起頻度、Jaccard 計数、Simpson 計数を用いて、地名と料理名に関する関連語ネットワークを構築した。各関連語ネットワークに含まれる関連語のうち上位下位関係にある関連語の割合を評価した。上位下位関係にある関連語を抽出するには Simpson 係数による関連度が適していることを示した。Simpson 係数による地名の関連語ネットワークでは、上位下位関係にある関連語が 88% の割合で抽出され、料理名の関連語ネットワークにおいては 56% の割合で抽出された。今後は、複数の属性の単語で構成される関連語ネットワークの構築と、関連語間の関連性の特定が課題となる。

文 献

- [1] Chen, Z., Liu, S., Wenyin, L., Pu, G., and Ma, W., Building a Web Thesaurus from Web Link Structure, in Proceedings of the ACM SIGIR, pp 48–55, 2003.
- [2] Fleischman, M., Hovy, E., and Echiabi, A., “Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked.”, In Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics, pp. 1–7, 2003.
- [3] Hearst, M. A., “Automatic acquisition of hyponyms from large text corpora.” In Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545, 1992.
- [4] 新里圭司, 鳥澤健太郎, HTML 文書からの単語間の上位下位関係の自動獲得, 情報処理学会研究報告 自然言語処理研究会報告, Vol.2003, No.108 pp. 95–102, 2003.
- [5] 荒牧英治, 今井健, 梶野正幸, 美代賢吾, 大江和彦, 自然言語処理によるオントロジー構築支援の試みとその問題点, 言語処理学会第 13 回年次大会 併設ワークショップ「言語的オントロジーの構築・連携・利用」, pp.51-54, 2007.
- [6] 安藤まや, 関根聡, 石崎俊, “定型表現を利用した新聞記事からの下位概念単語の自動抽出”, 情報処理学会研究報告 2003-NL-157, pp. 77–82, 2003.
- [7] 石田啓介, 松尾豊, 安田雪, ネットワーク作る君: Web マイニングを用いたネットワーク自動抽出システム, 第 21 回人工知能学会全国大会, 1B2-6, 2007.
- [8] 大澤幸生, ネルス E ベンソン, 谷内田雅彦, KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌, D-I, Vol.J82-D-I, No.2 pp. 391–400, 1999.
- [9] 榊剛史, 松尾豊, 石塚満, Web 上の情報を用いた関連語のソーラス構築について, 自然言語処理, Vol.14, No.2, pp.3-31, 2007.
- [10] 廣川佐千男, 下司義寛, 和多 太樹, 文書群からの概念グラフの構成, 情報処理学会研究報告 自然言語処理研究会報告, No.94, pp. 79–84, 2005.
- [11] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満, Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol.20, No. 1, pp. 46–56, 2005.
- [12] 吉田知訓, 山本啓司, 栗田恵理, 高野敦子, 北村泰彦, 半体系化 Web 情報源からの体系化知識ベースの構築, 情報処理学会第 69 回全国大会, 1N-3, 2007