Wikipediaからの概念抽出に基づくなぞかけ文の自動生成

内村 圭佑† 灘本 明代‡

†甲南大学理工学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1
‡甲南大学知能情報学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †math_55jp@yahoo.co.jp, ‡nadamoto@konan-u.ac.jp

あらまし 本研究では，Webページからなぞかけ文を自動生成することを目的としている．具体的には入力された単語(名詞)と得られた単語(名詞)との間にある意味的距離を，日本語 Wikipedia を用いた簡単な上位概念抽出手法にて測り，最終的にある程度意味的に離れた単語を用いてなぞかけ文を自動生成することを目指す.

キーワード なぞかけ，意味的距離，Wikipedia，上位概念抽出

Automatic Creation of NAZOKAKE, based on High Level Concept from Wikipedia

Keisuke UCHIMURA†, and Akiyo NADAMOTO‡

†Faculty of Science and Engineering 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan
‡Dept. of Intelligence and Informatics 8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo, 658-8501 Japan
E-mail: †math_55jp@yahoo.co.jp, ‡nadamoto@konan-u.ac.jp

Abstract We propose automatic creation of NAZOKAKE from Web. When the system creates NAZOKAKE automatically, we first compute the semantic distance between user's input word(noun) and the system output word(noun) by using the concept hierarchy which is extracted by Wikipedia. Next it extracts the words that the semantic is far from user's input word, and it creates NAZOKAKE by using the words.

Keyword NAZOKAKE, Meaning distance, Wikipedia, Extract upper concept

1. はじめに

近年のインターネットの普及に伴い，一般の人々が取得できる情報量は数年前と比較して格段に増加し，インターネットは情報の宝庫となっている．しかしながら，我々はこれらの情報の一部を利用するにすぎなく，これらを十分に使いこなしていないのが現状である．特に，現在の若者は語彙力・発想力に乏しいと言われている[1]が，我々はこの語彙力・発想力の強化にインターネットの情報を使えないかと考え，インターネット上にある様々な情報から発想を支援するシステムとしてなぞかけ検索を提案する.

「なぞかけ」とは「A とかけて B ととく．そのこころは C+D.」という定型句に，ある程度の関連性をもった単語を当てはめて楽しむ日本に古くからある言葉遊びの一種である．その起源は享保１３年（１７２８）に刊行された京都の俳人・重田梧山の著書『謎車氷室桜』（なぞぐるま・ひむろのさくら）[2]であり，また 686 年に天武天皇が問うたとされる無端事（あとなしこと)[3]であり，また平安時代に始まった芸能で新年を祝った万歳（まんざい）から派生した尾張万歳（おわりまんざい）や三河万歳（みかわまんざい）である[4]等とも言われている.

本研究ではこの「なぞかけ」の構造に注目し，単語 A と D を入力とし, A と D に各々関連する単語 B と単語 C を抽出することを試みる．例えば,「アイロンとかけてギャンブルととく．そのこころはカケル（かける/賭ける）とアツい.」の場合，入力を A:アイロン，D:アツい， とし，B:ギャンブルと C：カケルを抽出することを行う．ここで，単語 C と単語 D は単に A と B との関連語ではなく，A と C+D，B と C+D は関連があるが，A・B・(C＋D）のすべての関連は弱く，C+D は同音異字の２つのセットであるといえる．そこで本論文では，単語 A を名詞，単語 D を形容詞として，これら A，D を入力とし，A と D とに関連する同音異字の B と C の２つのセットを Web から取得することを行う．具体的には単語 B の抽出には，日本語 Wikipedia を用

Copyright ©2009 by IEICE

いた簡単な上位概念抽出手法を提案し，その抽出した概念構造を用いて単語間の意味的距離を測りBを抽出し，単語Cに関してはWeb検索の結果を用いる．

なぞかけには先の例のように「アイロンとかけてギャンブルととく．そのこころはカケル（かける/賭ける）とアツい．」といった同音異字のタイプと「３月とかけて１日ととく．そのこころは，月の前にサン(Sun/３)がくる．」といったトンチがはいっているタイプの 2 パターンが考えられる．本研究では，誰にでも分かりやすいなぞかけを生成することを目的とし，前者の同音異字のタイプを生成することを行う．

なぞかけ検索の特徴は以下の通りである．

- Wikipedia から単語の上位概念を抽出する．
- 抽出した上位概念を用いて単語間の意味的距離を求める．
- 意味的距離を用いて単語セットを抽出し，なぞかけ文を自動生成する．

以下，2 章では関連研究を，3 章では Wikipedia を用いた概念構造の抽出を，そして 4 章ではなぞかけの自動生成について述べ，5 章で実験・考察，6 章でまとめと今後の課題について述べる．

2. 関連研究

2.1. なぞかけ生成

前田ら[5]は，データベースを用いた面白いなぞかけの生成の研究を行っている．データベース内には単語が任意に収納されており，それらにはさらに付加情報としてある単語Aからある単語Bに対する関連度なども含まれている．前田らはアンケートによって得られた関連度も提案している．なぞかけとは言葉を用いた遊びの一種で，つまり言葉の意味や関連度というのはとても大事な要素の一つである．そのためなぞかけの生成においてそういった辞書の機能を果たすデータベースを作っておくことはとても有効であるといえる．しかしながら，それでは時事の単語に対応することが出来ないという短所もある．仮に対応しようとしてもその都度アンケートを取って関連度を測らなければならないため，あまり現実的であるとは言えない．そこで我々はなぞかけの生成手法として，人手により作成したデータベースを用いず，Web を用いて単語間の関連度をリアルタイムに自動で求める事を行う．これにより，前田らの研究のアンケートで得られるほどの正確な数値化の出来る関連度は期待できないが，時事の単語に対応することが可能となる．また，同じ入力であったとしてもその時期によって異なる結果を得ることのできる時代に即したなぞかけの生成が可能である．

2.2. Wikipedia からの Web オントロジー自動構築

中山ら[6]による研究では，Wikipedia から独自の手法を用いたオントロジー構築を提案している．彼らは Wikipedia 内の記事における文法に着目し，独自の手法として LSP 法(記事のリード部分を重要文と見做して解析する手法)と ISP 法(記事の中から重要な文章を抽出して解析する手法)を提案し，それらを用いたオントロジー構築を行っている．しかしここで扱われている記事というのはあくまで英語 Wikipedia のものであり，日本語 Wikipedia には対応していない．つまり LSP 法においては記事の中からまず冒頭文を抽出してくる必要があるが，ここでは英語の文法のしくみを考慮して，文章の初めの単語の頭文字が大文字であることやピリオド(もしくは改行コード)の直前が小文字であること等を利用し，単純なルールで効率よく抽出を実現している．我々の研究では日本語 Wikipedia を用いた上位概念抽出を行うので，これを日本語 Wikipedia に対応させようと考えたとき，上記の単純な冒頭文抽出のロジックは言語構造が違うので通用せず，英語 Wikipedia で構築されるオントロジー程の精度は望めないが，冒頭文は同じく is-a 関係にあることから LSP 法はかなり実用的な手法であると言える．また，中山らによると Wikipedia のリンク構造を用いた解析手法[7]は情報量の少なさが原因で，満足する結果が得られないことなども挙げられている．よって本研究においても Wikipedia 解析に際してはリンク構造を用いずに行うこととする．

3. Wikipedia を用いた概念構造の抽出

本章では Wikipedia からの簡単な上位概念抽出方法と，意味距離の計測方法を提案する．提案手法は，記事のリード部分(冒頭文)を重要文と見做して解析する手法，LSP(Lead Sentence Parsing)法[3]に基づき，手順を進める．

3.1. Wikipedia からの上位概念の抽出

Wikipedia 内の文章は多くの場合冒頭文が is-a 関係にあることに注目し，この関係を利用して単語の上位概念を抽出することを行う．具体的には，一文目の最後に出てきた名詞をその単語の上位概念とする．実際に抽出される上位概念は精度の低い上位概念抽出である場合がある．つまりはすぐ上にある概念ではなく，場合によってはある単語の上位概念の上位概念，もしくはそれ以上上の上位概念である場合があるが，本研究ではすぐ上の上位概念を必要としないため，本手法を用いる．実際には以下に示す精度の結果を得ることが出来る．

- [0] => マティーニ [1] => カクテル [2] => 飲料 [3] => もの
- [0] => 塩化ナトリウム [1] => 物質 [2] => 存在 [3] => 呼称
- [0] => 温暖化 [1] => 現象 [2] => 事実 [3] => 表現 [4] => 事

ここで[0]は初めに入力された単語を,[1]は[0]の,[2]は[1]の，[3]は[2]の上位概念を示す

Wikipedia を用いた上位概念抽出の場合，Wikipedia の特徴から，クエリに対して冒頭のパラグラフを抽出する．特徴とは，Wikipedia ではクエリに対してパラグラフの頭が「 クエリ」の形で出現するケースが多く，またパラグラフの終わりは「 」で表されるというものである．次に形態素解析器を用いて，抽出されたパラグラフの最初に出てくる「。」の手前に出現する名詞を上位概念として抽出し，次はそれをクエリとして同じ作業を繰り返す(図 1 参照)．

3.2.意味距離

単語 a_0，b_0 において得られた上位概念を用いて，2 つの単語の間にある意味的距離を以下の手順で求める．

1. a_0 の上位概念を 15 件まで取得する．(※1)
2. b_0 が a_n ($n = 0, \ldots, 15$)に一致しないことを確認し，これを b_{15} まで繰り返す．
3. 2.で b_m($m = 0, \ldots, 15$)が a_n に一致する時，その時点で 2 つの単語 a_0 と b_0 は意味的に近いとする．
4. 2.が最後 $a_n \neq b_{15}$ を確認すれば，a_0 と b_0 は意味的に離れているとする．

上記の手順で意味的な距離を求めるが，3.で説明している $a_n = b_m$ となったその際，a_0 と b_0 は a_n(= b_m)からのツリー構造として表わされるため数値としての距離は $n + m$ で求められる．

4. なぞかけの自動生成

この章ではなぞかけを自動生成するための仕組みを説明する．本論文では，同音異字を用いたなぞかけの自動生成を目標とする．同音異字を用いたなぞかけの例として「アイロンとかけてギャンブルととく．そのこころはカケル(かける/賭ける)とアツい.」を挙げ，これ(以下「甲」と略)を使ってこれよりシステムの説明をしていく．

初めに人がなぞかけを生成の手順を説明する．甲の場合まず，「アイロン」を足掛かりとして，それがどんなものかを考える．そこからアイロンとは「かけると

※1. WikipediaAPI において，1 分間に 30 回以上のクエリ送信は控えるように記載してある. 両単語の上位概念取得数が最大値の時，クエリ送信数が合わせて 30 回以内になるよう両単語 15 個以下取得に抑える. また上位概念が 15 個得られるケースは稀である為, 決して 15 個というのは少なくない.

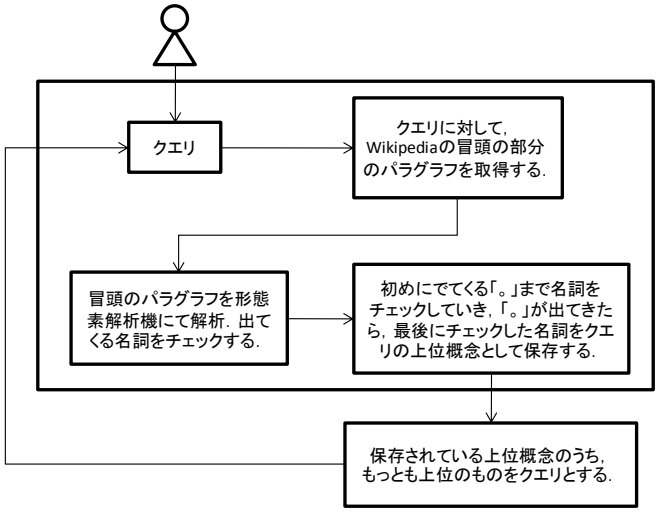

図 1. Wikipedia からの上位概念抽出

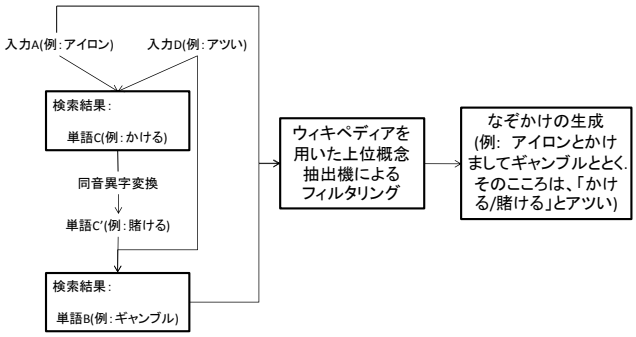

図 2. なぞかけ生成システムの流れ

アツい」ものであるということを連想し，次に，同じく「かけるとアツい」ものについて考える．ここで「かける」に注目して，これを同音異字である「賭ける」と置き換え，「賭けるとアツい」ものを考え出すことで一種の駄洒落要素を付け加える．これによりなぞかけをより面白いものとすることが出来る．そして最後に「賭けるとアツい」ものとして「ギャンブル」を連想すれば，それらを組み合わせて甲が出来る．

次に上記の作業をコンピュータにさせる手法を説明する．ユーザーから入力された「アイロン(名詞)」と「アツい(形容詞)」から，それら二つの単語を結びつける単語として「かける(動詞)」を得るために，「アイロン　アツい」をクエリとした時の Web の検索結果を用いる．この「かける」の抽出基準は，検索結果の中の単語(動詞)の頻出度により決定することとする．次に「かける」を同音異字変換し「賭ける」とする．これは人がなぞかけを生成する過程における，駄洒落要素を絡む部分に相当する．得られた「賭ける(動詞)」と，入力された単語「アツい(形容詞)」を用いて再度「賭ける　アツい」をクエリとした yahoo 検索を行い，検索結果から頻出度を基準として「ギャンブル(名詞)」

を抽出する．これをもって「アイロン」に対して，同じく「かけるとアツい」ものとして「ギャンブル」が選出され，それらの単語を組み合わせることで甲を生成する．

駄洒落要素はユーザーのセンスによって十人十色の組み合わせが予想される部分である為，一概にシステム側で選択を行うのは好ましくないと考える．よってあくまでシステムとしては同音異字を選択肢としてユーザーに示し，それらをユーザー側に選択させる方式をとる．つまり上記の例では「かける」の同音異字として「賭ける」を選択したが，ユーザーの好みで「掛ける」や「駆ける」等の単語を選択することが可能であるし，その場合最後に得られる単語も選択された単語に合うものに変わってくる．これによりユーザーは自分の嗜好に沿ったなぞかけの生成を目論むことが可能となる．

最後に，入力された単語「アイロン」及び，得られた単語「ギャンブル」に対して，Wikipediaに基づく上位概念抽出器を用いて単語間の意味的な距離を測る．それによって二つの単語が意味的に離れていることをチェックするためである．これは最終的に得られた単語が入力単語である「アイロン」と意味的に近かった場合，「アイロン」と「アイロンに意味的に近いもの」が，同じく「かけるとアツい」のは当たり前であり，つまりそれらを用いたなぞかけは面白みに欠けるからである(図2参照).

5. 実験・考察

ここではシステムの有用性を測る為に，予備実験を行う．実験は被験者1名を対象とし，無作為に単語を二つ(名詞(A)と形容詞(D))入力する．被験者は途中システムが出力した同音異字(C’)を選択肢の中から選ぶ．実験結果の表記例は以下の通りである。

- Aとかけて Bととく．そのこころは(C/C’)⇒D.

ここで C’はユーザーが選択した C の同音異字である．以下に実験結果の一部を表記例に従って示す.

- アイロンとかけて男ととく. そのこころは(かける/賭ける)⇒アツい.
- ベッドとかけて計画ととく．そのこころは(ねる/錬る)⇒眠い.
- カメラとかけて人ととく．そのこころは(とる/盗る)⇒綺麗.
- 思い出とかけて写真ととく. そのこころは(よみがえる/甦る)⇒古い.
- 部屋とかけてダイエットととく. そのこころは(めざす/目指す)⇒綺麗
- 宝くじとかけてプレゼントととく．そのこころは(あたる/当たる)⇒嬉しい.

上記のなぞかけの内，上の３つは成功例であるが，下の３つは失敗例であると言える．それは被験者が C’の部分にCと同じ意味の単語を選択しており，その結果なぞかけ自体に駄洒落の要素が含まれていない為である．つまりユーザーは同音異字を選択する際，なるべく C に対して意味的に離れた C’を選ばないと面白いなぞかけを得ることが出来ない．これを踏まえ，同音異字の選択肢は，あらかじめCに近い意味を持つものを取り除いた形でユーザーに示すべきであるという結論が得られた.

6. まとめと今後の課題

本研究では Wikipedia からの上位概念抽出と，それに基づくなぞかけの生成手法を提案し，また実験の結果から，当初の目的であった Web からのなぞかけ文自動生成を達成した．これによりインターネット上に存在する膨大な量の情報の中から，少しでも有用であり，ユーザーの発想を支援できる情報を提示できるなら重畳である．また今後の課題を以下に記す.

- 日本語 Wikipedia からのより精度の高い上位概念抽出手法を提案し，単語間の意味距離を厳密に測る.
- 実験で得た結果を活かして，ユーザーに与える同音異字の選択肢を絞り込む.
- 現段階でユーザーからの入力は名詞と形容詞に限られるが，それ以外の品詞にも対応したルールを考え，生成されるなぞかけの幅を広げる.

謝辞

本研究の一部は，平成 20 年度科研費特定領研究域「Web2.0 時代のコミュニティ型コンテンツのコンテンツホール検索に関する研究」(課題番号：19024072，代表：灘本明代)による．ここに記して謝意を表します.

文　献

[1] 教室を超えて生きる国語力 藤原正彦インタビュー
http://benesse.jp/berd/center/open/syo/view21/2005/04/s01toku_13.html

[2] 【お笑いメルマガ】なぞかけめ～る第 3 号[まぐまぐ!]
http://archive.mag2.com/0000159712/20050621200000000.html?page=7&js

[3] なぞかけユーモア（吉村外喜雄のなんだかんだ）
http://www.noevir-hk.co.jp/magazine/2008/06/post_471.html

[4] 元祖！俺は熱いうちに打て！：なぞかけ講座
http://shariki.hamazo.tv/c9613.html

[5] 前田実香，鬼沢武久，”言葉の関連性とおもしろさを取り入れたなぞかけ生成”，2004「エンタテイメント感性」ワークショップ講演論文集，pp.37-42, December, 2004

[6] 中山浩太郎，原隆浩，西尾章治郎，”自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジー自動構築”，日本データベース学会論文誌，Vol.7, No.1, pp.67-72, June, 2008

[7] 桜井慎弥，手島拓也，石川雅之，森田武史，和泉憲明，山口高平，”日本語 Wikipedia からの汎用オントロジーの構築と評価”，第 22 回人口知能学会全国大会論文集 3F2-3，(2008.6)