平成18年度

筑波大学第三学群情報学類

卒業研究論文

題目
日本語クロスワードパズルにおけるヒント文自動生成

主専攻　　情報科学主専攻

著者　　福富崇博

指導教員　山本幹雄

要　旨

近年、クロスワードの普及に伴って、クロスワードを自動的に解答・生成する研究に注目が集まっている。クロスワードを生成する手法は、大きく 2 つに分類することができる。盤面の白マス・黒マスの位置を決定し、白マス位置に単語を埋め込む工程と、キーワードにヒント文を付与する工程である。中でも日本語クロスワードを対象として生成されたヒント文はスリーヒント形式のみであり、まだまだ研究の余地は大きい。

本研究では専門雑誌に掲載されている日本語クロスワードからヒント文 984 文をランダムに収集・分類し、実際の日本語クロスワードでは説明文・虫食い文・連想形に属するヒント文が高い割合を占めることがわかった。本論文では辞書やコーパス、WWW 上に存在するオンラインニュース、ブログ等の言語資源を利用することで、クロスワードヒントの作成支援を検討する。

目 次

図 目 次

表目次

第1章　はじめに

1.1　研究の目的と背景

クロスワードパズル (以下、クロスワードと略記) は、1913 年当時ニューヨーク・ワールド紙日曜版「FUN」ページを担当していたアーサー・ウィンによって発明されたものといわれている。その後 1922 年には英国に紹介され、1925 年には日本でも「サンデー毎日」誌が連載をはじめた。今日、クロスワードは世界中に普及しているが、これらは各国の言語文化に依存しながら独自の進化をとげている [1]。中でも代表的な形式が米国、英国、日本で使われているもので、これらの形式のクロスワードをそれぞれ米式クロスワード、英国クロスワード、日本式クロスワードと呼ぶことにする。

クロスワードの例を図 1.1 に示した。クロスワードは、与えられる「ヒント」と呼ばれる文章 (以下では、クロスワードヒントと呼ぶ) から、単語 (以下、キーワードと呼ぶ) を推測し、タテヨコに交差した「グリッド」と呼ばれるマス目の空欄を埋めつくすパズルゲームである。クロスワードヒントからキーワードを推測する部分はある種の常識テスト、キーワードを盤面に埋め込む部分は探索問題となっており、これら全く異なった性質を持つ問題が組み合わさっている点が、クロスワードの最大の魅力であるといえるだろう。これは学術的な分野においても同様のことがいえ、例えば常識テストを機械的に解答・生成することは人間の連想や常識といった機能の実装であり、自然言語処理の分野においてたびたび研究の対象とされる [2][3][4][12]。また、盤面にキーワードを埋め込む部分、及び盤面の機械的生成は制約充足問題に定式化でき、これらは人工知能の分野において研究がされている。このようにクロスワードはパズルゲームとしてだけでなく、研究対象としてとても魅力的な素材である。以下ではクロスワードに関する研究を解答手法・生成手法の 2 つの視点から述べることにする。

クロスワードを解答する手法として代表的なものに、1999 年に発表された Proverb がある [2][3][4]。このシステムは米式クロスワードを自動的に解答するもので、「New York Times」に掲載されたクロスワードを対象に単語正解率[1]で 95.3%と極めて高い性能を示した。一方、日本式クロスワードの自動解答を目指したものには佐藤の研究がある [12]。こちらは単語正解率で 44%であり、Proverb と比較するとあまりよい成果とはいえない。これは 1)Proverb が総計 5142 面のクロスワードデータベースを利用して実現されているのに対し、日本式クロスワードにはそのような膨大数のデータベースが存在しないこと、2) 日本式クロスワードのグリッド制約[2]が米式クロスワードと比較して弱いことによるものと思われる。

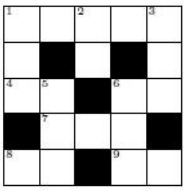
[1] クロスワードに含まれるキーワードのうち、いくつ正解したかの割合

[2] クロスワードにおいてタテヨコの交差は埋めるべきキーワード候補を制限する働きをもち、これをグリッド制約と呼ぶ。

クロスワードを生成する手法は、大きく 2 つに分類することができる。1 つはクロスワード盤面の白マス・黒マスの位置を決定し、白マス位置に単語を埋め込む研究 [5][8]、もう 1 つはクロスワードヒントを機械的に生成する研究 [9] である。加えてキーワードは一つのテーマに基づいて選ぶのが主流のため、関連語を収集する研究も存在する [6]。しかしながらクロスワードを生成する研究は、どの分類においても少ない。特に日本式クロスワードを対象に生成されたクロスワードヒントはスリーヒントクロスワードのみであり、まだまだ研究の余地は大きいといえる。

クロスワードヒントを機械的に生成しようとしたとき、まずはじめにかんがえられるのは大量のクロスワードヒントデータベースからキーワードに対応するクロスワードヒントを抽出することである。しかしながら前述のように、この手法を実現できるほど膨大な日本式クロスワードのデータベースは存在しない。本論文では、辞書やコーパス、WWW 上に存在するネットニュース、blog 等の言語資源を利用することで、クロスワードヒントの自動生成を検討した。

これより第 2 章で日本式クロスワードの特徴を述べ、第 3 章にて実際の日本式クロスワードを収集・分析する。続いて第 4 章で検討した生成手法について説明し、第 5 章ではそれぞれの生成手法を実験・評価する。第 6 章で、本論文のまとめと今後の課題を述べる。



ヨコ

1 マイクロソフト社の会長 (5)
4 ○○は友を呼ぶ (2)
6 ⇔借り (2)
7 西インド諸島の別名. ○○○諸島 (3)
8 2006 年 10 月 9 日, 北朝鮮が「○○実験を実施した」と発表 (2)
9 この式の○○を求めよ (2)

タテ

1 ○○○をかけあって優勝を祝った (3)
2 モンゴル高原で見られる移動式住居 (2)
3 日本ハムの優勝に貢献した新庄○○○選手 (3)
5 北朝鮮軍兵士 5 人が一時軍事境界線から侵入、韓国軍兵士が○○○射撃 (3)
6 株式の価格 (3)

図 1.1: クロスワードの例

第2章　クロスワードパズルの特徴と自動生成の流れ

本章では代表的なクロスワードである米式クロスワード、英式クロスワード、日本式クロスワードについて概説し、続いて日本式クロスワード自動生成の流れを説明する。

2.1　クロスワードパズルの特徴

クロスワードは世界中に普及し、各国の言語文化等に依存しながらそれぞれ独自の進化をとげている。現在、クロスワードの形式は実に多種多様に存在するが、中でも代表的なものに米式クロスワード、英式クロスワード、日本式クロスワードがある。本節では、それぞれのクロスワードを作成する際の基本ルールと、そこから導出される特徴を概説する。

2.1.1　米式クロスワードパズル

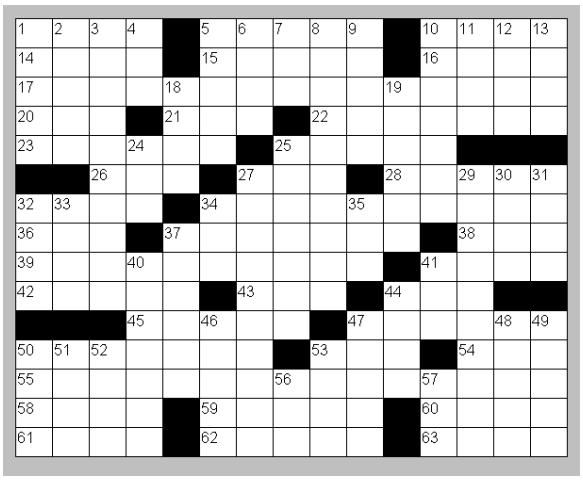


図 2.1: 米式クロスワードの盤面

米式クロスワードの盤面を図 2.1 に示した[1]。米式クロスワードを作成する際の基本ルールには次のものがあり、主要な新聞に掲載されるようなパズルでは厳密に守られている。

[1] http://en.wikipedia.org/wiki/Image:American_crossword.png

A1. 黒マスは対角線に対して対象に配置する。

A2. キーワードには 3 文字以上の単語を使用する。

A3. すべての文字は、タテヨコ両方のキーワードの一部となっている。

A4. 白マスはすべて連結している。

A5. 黒マスの連続に制限がない。

A6. 黒マスは盤面全体の $\frac{1}{6}$ 以下である。

クロスワードにおいてタテヨコの交差は埋めるべきキーワード候補を制限する働きをもち、これをグリッド制約と呼ぶ。米式クロスワードは盤面のすべての文字はタテヨコ両方のキーワードの一部でなければならない (A3) ため、グリッド制約が強いのが特徴である。黒マスが盤面全体の $\frac{1}{6}$ 以下でなくてはならない (A6) というルールがあるが、タテヨコ両方のキーワードの一部である条件から、黒マスの数が少なくなるのは当然である。図 2.1 のように白マス数が多いのもまた、米式クロスワードの特徴のひとつである。米式クロスワードを作成する際のルールには、Web サイト「CrossDown」[2]が詳しい。

2.1.2 英式クロスワードパズル

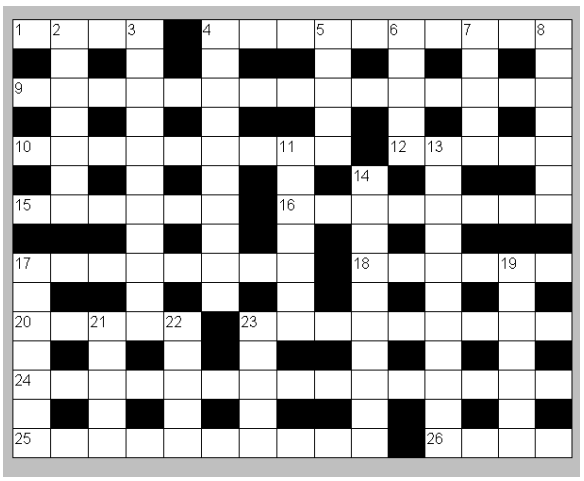


図 2.2: 英式クロスワードの盤面

英式クロスワードの盤面を図 2.2 に示す[3]。英式クロスワードを作成する際の基本ルールには、次のものがある。

[2]http://www.crossdown.com/howtomake.htm
[3]http://en.wikipedia.org/wiki/Image:British_crossword.png

B1. 黒マスは対角線に対して対称に配置する。

B2. キーワードには 3 文字以上の単語を使用する。

B3. すべての文字は、タテヨコどちらか、または両方のキーワードの一部となっている。

B4. 白マスはすべて連結している。

B5. 黒マスの連続に制限がない。

B6. 黒マス数に制限がない。

英式クロスワードは盤面の文字はタテヨコどちらか一方のキーワードの一部であればよい (B3) ため、グリッド制約が弱い。また、黒マス数においても米式クロスワードとは対称的で、制限が設けられていない (B5)。第 2 章では、クロスワードが常識テストと探索問題の組合せで構成されることを述べた。キーワードを推測する手段として、米式クロスワードは探索問題の占める役割が大きく、英式クロスワードは常識テストの占める割合が大きいといえるだろう。

2.1.3 日本式クロスワードパズル

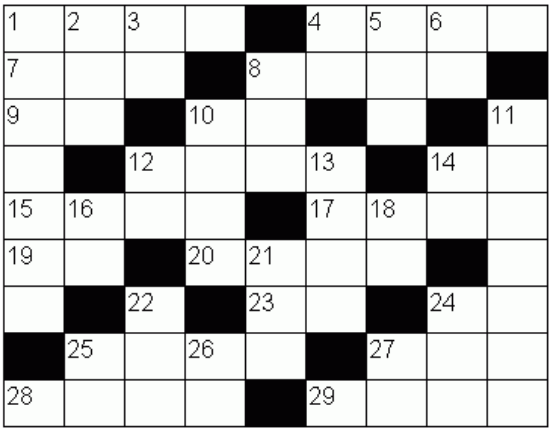


図 2.3: 日本式クロスワードの盤面

日本式クロスワードの盤面を図 2.3 に示す。日本式クロスワードを作成する際の基本ルールには、次のものがある。ただし前述した米式クロスワードや英式クロスワードと比べると、これらのルールはそれほど厳密に守られていないようだ[4]。

[4]「絵ヒントクロス」や「タイポグラフィックスクロス」を代表に、形式にとらわれない多種多様なクロスワードが存在する。

J1. キーワードには 2 文字以上の単語を使用する。

J2. すべての文字は、タテヨコどちらか、または両方のキーワードの一部となっている。

J3. 白マスはすべて連結している。

J4. 黒マスをタテヨコに連続させてはならない。

J5. 盤面の 4 隅は白マスでなければならない。

J6. 黒マス数に制限がない。

日本式クロスワードは盤面の文字はタテヨコどちらか一方のキーワードの一部であればよい (J2) ため、グリッド制約が弱い。また、黒マス数には制限が設けられていないが、黒マスに関する明確な決まり事はいくつか存在するようだ (J4)(J5)。黒マスの占める割合が大きくなると、それだけ白マスの連結が多くみられるようになる。それに伴って、キーワードを推測の際、クロスワードヒントに期待されるキーワードを絞りこむための情報も高まる。日本式クロスワードにおいてクロスワードヒントの占める役割は、それなりに大きいといえるだろう。本研究では、日本式クロスワードを対象とする。これより特別な断りがない限り、クロスワードという表記は日本式クロスワードを指すものとする。

2.2 日本式クロスワードパズル自動生成の流れと本研究の位置づけ

2.2.1 クロスワードパズル自動生成システムの概要

前節では、日本式クロスワードとその他代表的なクロスワード形式とを比較し、日本式クロスワードの特徴を概説した。本節では日本式クロスワードの自動生成システムを定義し、システム内における本研究の位置づけを示す。近年のクロスワードにおいては、キーワードは 1 つのテーマに基づいて選ぶのが主流であるため、本システムでもテーマに基づいたクロスワードを生成するものとする。

テーマに基づいたクロスワード自動生成システムの概要を、図 2.4 に示した。本システムにおけるクロスワード自動生成の工程はテーマの決定・関連語収集・盤面自動生成・クロスワードヒント生成の 4 つに分けることができる。これより各工程で行われる操作について、詳しく説明する。

1. テーマの決定

盤面の単語を選ぶ基準となる、クロスワードのテーマを決める。本システムでは、クロスワードのテーマはユーザが入力する。

2. 関連語収集

前工程で決定したテーマに基づき、関連語を収集する。この工程の先行研究としては、WordNet からキーワード候補を抽出した研究がある [6] ものの、日本式クロスワードを

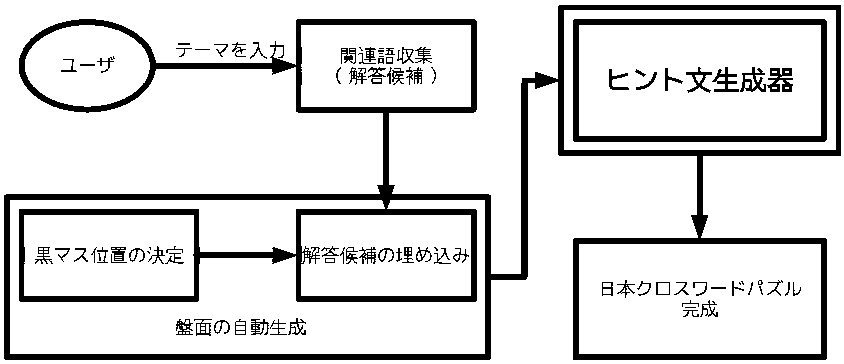


図 2.4: 自動生成システム

対象とした研究はみあたらない。しかしながら、日本語を対象に関連語収集を行った研究は多い [10][11]。

3. 盤面自動生成

クロスワード盤面の生成は、大きく 2 つに分けられる。1 つは黒マスの位置を決定する工程、もう 1 つは決定した空欄にキーワードを埋め込む工程である。本工程ではまず、黒マスの位置を決定する。黒マス位置の決定は、(2.1) 節で説明したクロスワード形式の基本ルールに基づかなければならない。日本式クロスワードの基本ルールには、次のものがある。

- キーワードには 2 文字以上の単語を使用する。
- すべての文字は、タテヨコどちらか、または両方のキーワードの一部となっている。
- 白マスはすべて連結している。
- 黒マスをタテヨコに連続させてはならない。
- 盤面の 4 隅は白マスでなければならない。

続いて、決定した空欄に単語を埋め込む。この工程は制約充足問題として定式化することができ、人工知能の分野において研究が盛んである [5][8]。

4. クロスワードヒント生成

決定したキーワードに、クロスワードヒントを付与する。クロスワードヒントを生成する研究には、スリーヒントクロスワードパズルの解作成システム [9] がある。このシステムのクロスワードヒント抽出部では、各単語の属性をクロスワードヒントの候補群としている。例えば「林檎」という単語に対応するクロスワードヒントの候補として、「果実」「花」「高木」等の単語が選出される。

2.2.2 本研究の位置づけ

日本式クロスワードは、盤面の文字はタテヨコどちらか一方のキーワードの一部であればよいため、グリッド制約が弱い。キーワードを推測する手がかりにはクロスワードヒントによるものとグリッド制約によるものとがあるが、日本式クロスワードではクロスワードヒントの手がかりが占める役割が大きいといえるだろう。今日、日本式クロスワードには多種多様なクロスワードヒントが用いられている。クロスワードヒントを生成する研究としてスリーヒントクロスワードの解作成システムがあるが、ここで作成されたクロスワードヒントは、日本式クロスワードで用いられているクロスワードヒントの中の、ほんの一部にすぎない。

本研究ではより多様なクロスワードヒントの生成手法を検討することにより、本格的に日本式クロスワードの作成を支援できるシステムの実現を目指す。

第3章　クロスワードの分析

これまで日本式クロスワードの特徴を概説し、日本式クロスワード自動生成システムにおける本研究の位置づけを示した。本研究の目的は日本式クロスワードにおける多様なクロスワードヒントの生成である。それでは、日本式クロスワードに用いられるヒントとはどのようなものなのか。本節では、クロスワード専門雑誌に掲載されている日本式クロスワードからヒントデータベースを作成し、日本語クロスワードに用いられるヒントの性質を検討する。

3.1　日本式クロスワードの収集

3.1.1　日本式クロスワードの収集

表 3.1: 収集クロスワード一覧

パズル ID	掲載元雑誌名	問題番号	価格	出版社
01	クロスワード Day 6 月号	Q18	420 円	双葉社
02	クロスワード Day 7 月号	Q1	420 円	双葉社
03	クロスワード Day 7 月号	Q8	420 円	双葉社
04	クロスワードキング 8 月号	Q41	420 円	インフォレスト
05	クロスワードキング 10 月号	Q30	420 円	インフォレスト
06	アロークロス 8 月号	Q5	420 円	インフォレスト
07	クロスワードメイト 10 月号	Q3	420 円	マガジン・マガジン
08	クロスワードメイト 10 月号	Q4	420 円	マガジン・マガジン
09	難問クロス vol.2	Q32	480 円	マガジン・マガジン
10	クロスワードランド 9 月号	Q38	420 円	白夜書房
11	クロスワードランド 11 月号	Q37	420 円	白夜書房
12	クロスワードハウス 8 月号	Q3	420 円	廣済堂出版
13	クロスワードセブン vol.1	Q37	420 円	桃園書房
14	クロスワールド 秋号	Q21	420 円	日本エディターズ

現在流通しているクロスワード雑誌 12 誌に掲載されていたクロスワードから、ランダムに 14 面のパズルを収集した。今回収集したクロスワードの情報を表 3.1 に示す。

日本式クロスワードにはヒントを文章ではなく絵で表した「絵ヒントクロス」や各ヒントを文字で絵を描いて表した「タイポグラフィックスクロス」など、テキストによらない特殊なクロスワードが存在するが、本研究ではこのようなクロスワードを収集対象から除外した。

3.1.2　クロスワードヒントデータベースの作成

(3.1.1) 節で収集したクロスワードからキーワードとクロスワードヒントのペアを抜き出し、ヒントデータベースを作成した。収集対象のクロスワード総数は 14 面で、抽出されたクロスワードヒントは 984 個であった。データベースの仕様は以下の通りである。

- 収録クロスワードヒント数:984 個
- クロスワードヒント・キーワード (カタカナ)・キーワード (漢字) をタブ区切りで記述
- 「%」から始まる行はコメント行

完成したクロスワードヒントデータベースの例を表 3.2 に示す。

表 3.2: 作成したクロスワードヒントデータベースの例

%クロスワード Day,Q1		
オスには立派な角がある、子どもたちに大人気の昆虫	カブトムシ	カブトムシ
多くの荷物を運ぶ自動車	トラック	トラック
玄関に取り付けているベル	リン	リン
ビルや橋の建設を行うこと	セコウ	施工
コーヒー豆を細かく挽いたもの	コナ	粉
⋮	⋮	⋮
さーっと降ってきて、さっとやむ雨	トオリアメ	通り雨
%クロスワード Day,Q8		
マイクを握って盛り上がろう。脳の血流がよくなるよ	カラオケ	カラオケ
俳句や短歌の総称。右脳と左脳がフル回転	ワカ	和歌
⋮	⋮	⋮

3.2　日本式クロスワードの分析

表 3.3: クロスワードヒントの例

	クロスワードヒント	キーワード
A	商品の値段。	価格 (カカク)
B	ナイル、アマゾン、信濃	川 (カワ)
C	日の当たる場所で○○○○○○。ポカポカしていい気持ち。	日向ぼっこ (ヒナタボッコ)

表 3.3 に、クロスワードヒントの例を示す。日本式クロスワードのクロスワードヒントには、様々性質が存在する。まずわかりやすい違いとして C が虫食いであることがあげられるが、この他にも形式的な違いとして、例えば A は 1 文、B は単語の羅列、C は 2 文によって

構成されることがわかる。他方クロスワードヒントの内容に着目すると、A はキーワードの説明文である。また、B ではキーワードに関連の深い単語を列挙することでキーワードを連想させており、C では「日の当たる場所」「ポカポカして」といったキーワードに関連の深い箇所を文中に含めることで、キーワードを連想させている。

本節では (3.1) 節で作成したヒントデータベースを用いて、日本式クロスワードヒントの持つ性質について分析する。

ここではクロスワード全体およびヒントに関する表層上の性質を分析する。(3.1) 節で作成したクロスワードヒントデータベースを対象に調査した結果は、表 3.4 のようになった。表 3.4 のパズル ID は、それぞれ表 3.1 のパズル ID と対応している。

表 3.4 中のデータは、左からそれぞれパズル ID・盤面の形[1]・盤面に含まれる全てのマス (白マス+黒マス) 中に黒マスが含まれる割合・キーワードの総数・キーワードの平均文字数・クロスワードヒントを構成する平均文数・クロスワードヒントを構成する平均単語数を表している。調査の際、クロスワードヒントを構成する文数として句点の出現回数を数え、クロスワードヒントを構成する単語数をカウントする際の単語の単位としては形態素解析システム「茶筌」[2]が切り出した形態素を使用した[3]。

盤面全体に占める黒マス割合の平均は、23.85%であった。(2.1) 節で述べた米式クロスワード作成時の基本ルールでは、米式クロスワードの盤面全体に占める黒マス割合は $\frac{1}{6}(16.67\%)$ 以下でなければならない。クロスワードにおいて黒マス数の占める割合が高くなると、それに伴ってグリッド制約が弱くなる。今回調査したクロスワード中において、盤面全体に占める黒マス割合が 16.67%を下回るものはみられなかった。

クロスワードヒントを構成する平均単語数は 9.17 単語であった。この結果は米式クロスワードの平均単語数が 2.5 単語 [7] であるのと比較して、極めて高い数字である。クロスワードヒントを構成する単語数の増加に伴って、クロスワードヒントの保持する、キーワードを推測するための情報も増加する。ゆえに日本式クロスワードではキーワードを推測する際に、クロスワードヒントの担う役割が大きいといえる。

クロスワードヒントを構成する平均文数は 1.13 文であった。日本式クロスワードで用いられるヒントは、ほぼ 1 文で構成されているといってよい。これより、クロスワードヒントを構成する 1 文のことをヒント文と定義する。本研究においてクロスワードヒントはすべて 1 文で生成するものとし、これをヒント文生成と呼ぶ。

[1] それぞれ、(縦マス数) × (横マス数) の形式で記述してある。盤面の形が長方形でないものは「変形」と記述し、括弧内に盤面に含まれるマスの総数を記述した。

[2] http://chasen.naist.jp/hiki/ChaSen/

[3] このとき句読点は除外し、○の連続 (虫食い) はそれをもって 1 単語とした。

表 3.4: 収集クロスワードの考察

パズル ID	盤面の形	黒マス数/マス数	キーワード数	キーワードの文字数	ヒントの文数	ヒントの単語数
01	11 × 11	21.49%(26/121)	51	3.43	1.06	7.43
02	13 × 13	23.67%(40/169)	78	3.05	1.01	8.51
03	13 × 13	23.08%(39/169)	73	3.23	1.36	11.42
04	14 × 14	23.47%(46/196)	85	3.21	1.08	8.91
05	変形 (196)	18.88%(37/196)	83	3.11	1.23	11.51
06	変形 (169)	40.83%(69/169)	96	3.23	1.02	4.38
07	13 × 13	23.67%(40/169)	83	2.93	1.05	10.19
08	13 × 13	22.49%(38/169)	79	3.09	1.15	10.65
09	14 × 14	24.49%(48/196)	81	3.20	1.32	12.95
10	13 × 13	21.89%(37/169)	79	3.14	1.08	7.35
11	12 × 12	29.75%(36/121)	61	3.26	1.23	11.16
12	8 × 8	18.75%(12/64)	28	3.34	1.17	9.17
13	12 × 12	20.83%(30/144)	59	3.56	1.0	5.69
14	11 × 11	20.66%(25/121)	51	3.43	1.06	9.08
計	-	-	987	-	-	-
平均	-	23.85%	-	3.23	1.13	9.17

3.3 クロスワードヒントの分類

本節では日本式クロスワードのキーワードを推測させる性質に着目した分類体系を検討・作成し、作成した分類体系に基づいてクロスワードヒントデータベースの分類を行う。クロスワードヒントは複数の文で構成されることもあるため、必ずしも 1 つのクロスワードヒントが 1 つの連想タイプに分類されるとは限らない。例えば「おさえとどめること。インフレを○○○○する。」というクロスワードヒント[4]において、「おさえとどめること。」部は「説明文」に分類され、「インフレを○○○○する。」部は「一般文章による虫食い」に分類される。分類結果を表 3.5 に示した。表 3.5 中のデータは、左からそれぞれ分類番号・分類タイプ・分類されたクロスワードヒントの例・分類されたクロスワードヒントの数を表している。

各分類タイプに分類されたクロスワードヒントの数は「説明文」が最も多く、続いて「虫食い文」、「連想形」が多い。これら 3 タイプを足し合わせた総数は 679 文と、全体のおよそ 7 割を超えた。本研究では、これら説明文・虫食い文・連想形のクロスワードヒントを生成の対象とする。

[4]対応する単語は「抑制」

表 3.5: ヒントタイプ分類表

分類番号	分類タイプ	クロスワードヒントの例	数
1	説明文	火を消すこと - 消火 (シヨウカ)	364
2	具体例	タンスやテーブルなど - 家具 (カグ９)	34
3	同意語	ぶどう酒のこと - ワイン	49
4	対訳	昼食- ランチ	49
5	反義語	⇔帰り - 行き (ユキ)	12
6	虫食い		283
6.1	慣用表現の一部が虫食い	○○も実力のうち - 運 (ウン)	51
6.2	一般文章の一部が虫食い	○○をゴクゴク鳴らして水を飲む - 喉 (ノド)	121
6.3	複合語の一部が虫食い	同じクラスの○○○○○生 - 同級 (ドウキユウ)	111
7	相対的な関係で説明	夫からみた、妻のお父さん -義父 (ギフ)	37
8	連想形		167
8.1	単語の羅列	ナイル、アマゾン、信濃 - 川 (カワ)	14
8.2	関連する文章	脳に「DNA」は関係ない - 遺伝子 (イデンシ)	153
9	その他		29
9.1	擬音語・擬態語・台詞	キャ〜！ヒィ〜！ - 悲鳴 (ヒメイ)	17
9.2	難読漢字	漢字で「鮪」- 鮪 (マグロ)	5
9.3	クイズ・なぞなぞ	森-木=？ - 林 (ハヤシ)	7

第4章　クロスワードヒントの生成

4.1　説明文の生成

説明文タイプのヒント文を以下に示す。

A.　商品の値段。　価格 (カカク)

(3.2) 節で行ったクロスワードヒントの分類では、説明文タイプのヒント文が最も多かった。説明文が得られる言語資源として、まず考えられるのは辞書である。本節では、辞書からヒント文を抽出する手法を説明する。

4.1.1　国語辞典を用いた説明文の生成

表 4.1: 学研現代新国語辞典データの例

アーケード	アーケード	〔洋風の大きな建物で〕柱列によって支えられた …
アース	アース	《名・他サ》感度をよくしたり、感電を防いだり …
アーチ	アーチ	大きな重みがかかるようなとき、上部を弓形にし …
⋮	⋮	⋮
もの-わすれ	モノワスレ	《自サ》〔年をとったりして〕記憶力が弱まるこ …
もの-わらい	モノワライ	（ーわらひ）他の人々からあざけり笑われること …
⋮	⋮	⋮

本研究では、説明文タイプのヒント文抽出に学研現代新国語辞典 (以下、辞書と略記) を用いた。辞書から抽出したデータ例を表 4.1 に示す。これより、辞書データに記述されている内容を左から順に見出語・読み仮名・コンテンツと定義する。それぞれの内容はタブで区切った。

以下、各工程での処理を説明する。

1. 見出語検索

   与えられたキーワードと一致する見出語を検索する。

2. 文単位に分割

手順 1 で検索した見出語のコンテンツを、文単位に区切る。文と文との境界を区別する基準には句点を用いたが、この例外を 2 つ設けた。例外とは、次のようなものである。「...」は任意の文字の連続を表す。

- 辞書特有の「...。また、...。」という表現は、この 2 文で 1 単位とする。
- 「... ー...」のように「」内にハイフンを含むものは句点を含まないが、これを「」1 組で 1 単位とする。

3. 不要部のフィルタリング

コンテンツ内には、次のような記述が混在している。

- 品詞タグ

  後に続く説明が、見出語がどの品詞の働きで用いられたものかを、タグで記述したもの。例えば「《名・他サ》感度をよくしたり、感電を防いだり...」で、「《名・他サ》」は後に続く説明が名詞、またはサ変他動詞として用いられたときのものであることを示している。

- 〔〕、〈〉、() 等、括弧で囲まれた文

  括弧の中で、後に続く説明内容を補足するもの。例えば「〔洋風の大きな建て物で〕柱列によって支えられ...」の説明文では、〔〕内部の記述によって後に続く文章が洋風の大きな建て物を前提とした説明であることを補足している。このような記述の処理には

  処理 1. そのままヒント文に利用する

  処理 2. 括弧を除去して、括弧内の記述はそのままヒント文の一部に利用する

  処理 3. 括弧内の記述はすべて除去する

  の 3 つの方法が考えられる。例えば「〔洋風の大きな建て物で〕柱列によって支えられた丸い天井をもつ通路。」の説明文に上記 3 つの処理を行うと、

  出力 1. 「〔洋風の大きな建て物で〕柱列によって支えられた丸い天井をもつ通路。」

  出力 2. 「洋風の大きな建て物で柱列によって支えられた丸い天井をもつ通路。」

  出力 3. 「柱列によって支えられた丸い天井をもつ通路。」

  となる。　クロスワードのヒント文として括弧内で説明を補足するのは不適切であるため[1]、処理 1 は処理候補から除外した。次に処理 2 と処理 3 であるが、処理 2 は処理 3 に比べて括弧内で補足された内容が含まれるため、キーワードを推測するための情報が豊富である。しかしながら処理 3 の出力中にも、キーワードを推測するための情報は十分に含まれており、処理 2 のように括弧内で補足された内容まで用いるのは、ヒント文として冗長である。本研究では、括弧内の記述はすべて除去することにした。

[1]収集クロスワードヒント 984 文中、このような形式のものはみられなかった。

- 文・単語
  - 見出語を含む文・単語

    クイズ問題中にクイズの答えが含まれるのが不適切であるのと同様に、ヒント文中にキーワードを含むものはヒント文として不適切である。見出語を含む文の処理方法を検討した。見出語を含む文には、次のようなものがある。

    例 1. 留守番　-　留守 (ルス)

    例 2. お経　-　経 (キョウ)

    例 3. 今日と同じ日付の日　-　今日 (キョウ)

    このような記述の処理には、

    処理 1. 見出語部分を虫食いにする 。

    処理 2. 見出語を含む文はヒント文候補から除外する。

    の 2 つの方法が考えられる。　見出語を含むヒント文候補は、複合語として頻繁に現れた。見出語部分を虫食いにすると想定した場合、例えばヒント 1 の「留守番」は「○○番」となる。また、ヒント 2 において「お経」は「お○○○」である。見出語を用いた複合語によって見出語と同じ意味を表しているゆえに、見出語以外の部分に含まれる見出語を連想させる情報を期待することはできない。これらの文を虫食いにしたものは、単語を連想させるための情報が圧倒的に足りない。また、ヒント文 3 の「今日と同じ日付の日」は「○○○と同じ日付の日」となる。これも見出語を含んだ文章を用いて見出語と同じ意味を表しているゆえに、見出語以外の部分に見出語に関する情報を期待することができない。このように、見出語を含んだ文についても、虫食いとしてヒント文に用いるのは不適切である。

    よって見出語を含むものは全て、ヒント文候補から除外することにした。
  - 見出語を含む文・単語

    見出語を含まない文とは、見出語でない単語のみを用いた見出語の説明文のことである。例えば、次のようなものがある。

    例 1. 「夕焼け小焼けー天気になあれ」　-　アシタ

    例 2. 終業。　-　卒業 (ソツギョウ)

    例 3. ある事柄について筋道を立てて自分の意見を述べしるした文章。　-　論文 (ロンブン)

  本手法では、これら見出語を含まない文をヒント文候補として用いる。

4. ヒント文出力

手順 3 までで抽出したヒント文候補をヒント文として適切な形に加工し、出力する。具体的なヒント文の形式には次のようなものがある。

- 虫食い文

「... ー...」の形式で記述された見出語を含む複合語、または文章を用いる。ハイフン部に見出語が入る。このハイフン部を○に置き換えることによって、虫食い文に加工する。○の数は、見出語の読みの文字数と一致させる。

- その他ヒント文候補　ヒント文候補を茶筌にかけ、ヒント文を構成する単語数を数える。単語数が 1 のものを複数個並べて、単語の羅列による連想形ヒント文として出力する。単語数が 1 より大きいものを説明文として出力する。

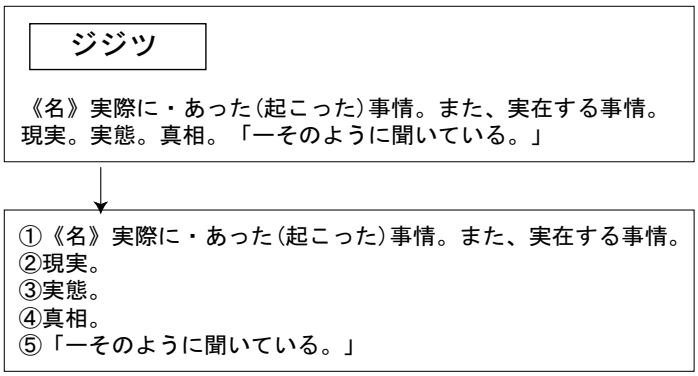


図 4.1: コンテンツの文切り

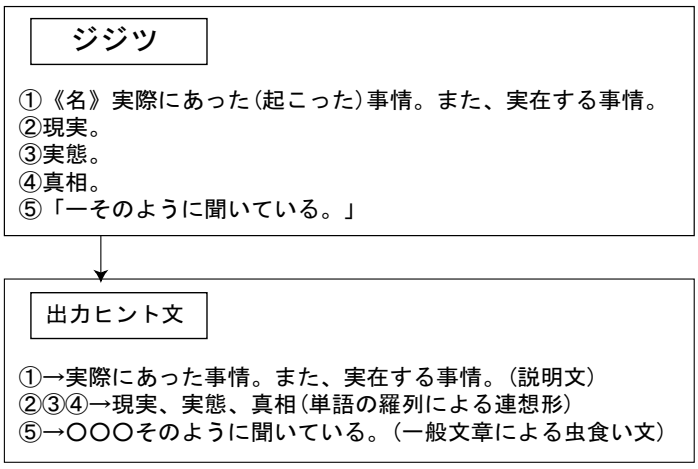


図 4.2: コンテンツ文の加工

4.2 虫食い文の生成

4.2.1 ニュース記事のタイトルを用いた虫食い文生成

表 4.2: ウィキニュースから取得したデータ例

（海外）台湾の淑珍・陳水扁総統夫人ら国防機密費の流用容疑にて起訴される
1100g の男児の心臓手術に成功-長野
12 月から速度規制解除-羽越線脱線事故
18 才タレントと飲酒した社員ら処分_フジテレビ
18 歳タレント飲酒にケツメイシ同席
⋮

一般にニュース記事のタイトルは 1 文で記事の内容を要約したものであり、タイトル中に含まれる単語同士が深く関わりを持つ。例えば「第 85 回天皇杯サッカー・浦和レッズ 25 年ぶり優勝」という記事では、「天皇杯」、「サッカー」、「浦和レッズ」、「優勝」の間に深い関連があるのがみてとれる。ヒント文はキーワードを推測するための手がかりである。一方で、ニュース記事のタイトルはそれぞれの単語が深い関連を持つため、ある単語を虫食いとしたとき、タイトル中のその他の単語にはキーワードを推測させるための手掛かりが豊富である。すなわち、ニュース記事のタイトルは虫食いヒント文を作成するのに非常に有用な資源であるといえる。

ウィキニュースのデータベースから取得したニュース記事のタイトルを、表 4.2 に示す[2]。

記事タイトルから虫食い文を生成する流れを図 4.3 に示した。

ニュース記事から虫食いヒント文を生成する手法を説明する。本手法ではまず、1) ニュース記事のタイトルを単語に分割し、それぞれの読み仮名・品詞情報を得る。続いて 2) 工程 1 で取得した品詞情報に基づき、数・地域-一般以外の名詞を読み仮名文字数の”○”に置換する。この際単語分割及び単語の読み仮名・品詞情報の取得には、形態素解析システム「茶筌」を用いる。生成したヒント文はキーワードとペアの形でデータベースに保存し、後からキーワードにヒント文を付与する際はデータベースからキーワードを検索する。

4.2.2 複合語による虫食い文

複合語による虫食い文を以下に示す。

例.　スポット○○○を浴びる　　ライト (ライト)

複合語による虫食い文とは、キーワードが複合語を構成する単語の 1 つとなる虫食い文である。上の例ではキーワード「ライト」が複合語のスポットライトを構成する単語の 1 つで

[2]http://download.wikimedia.org/jawikinews/20061222/

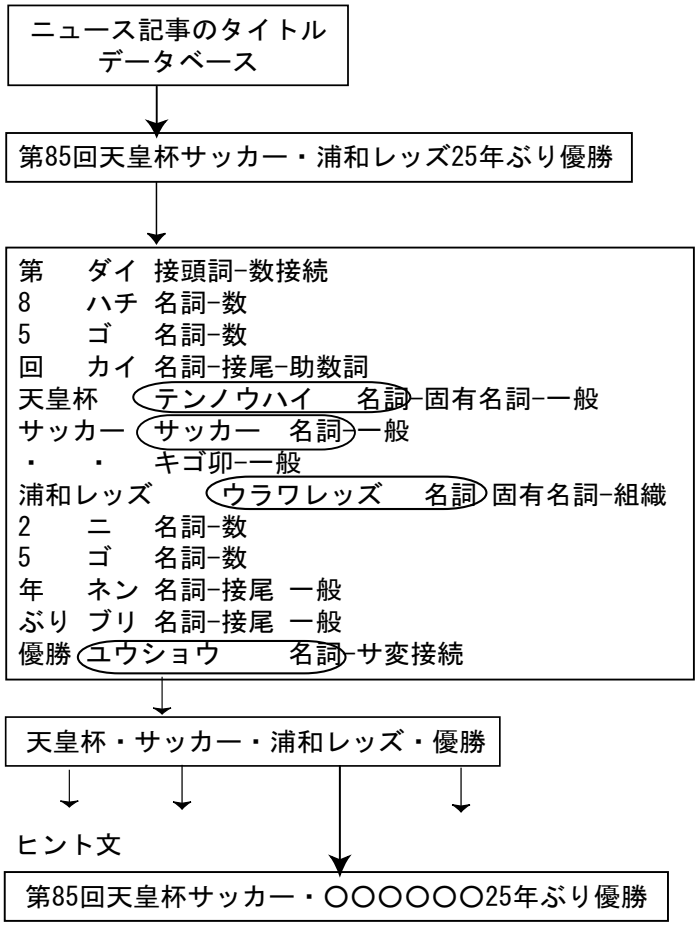


図 4.3: ニュース記事タイトルを用いたヒント文抽出

あることにより、「スポット〇〇〇」がキーワードを推測するための手掛かりとなる。がキーワードを推測するための手掛かりとなる。また、「浴びる」というスポットライトと関連の深い文脈を加えることによって、さらにキーワードはさらに絞りこまれる。

このように、複合語による虫食文は

1. キーワードが複合語の一部であること
2. 複合語と関連の深い文章であること

の 2 段構えでキーワードを推測させている。本手法では複合語を含む虫食い文を生成する手順としてまず複合語を、続いて複合語を含む文を抽出する。

本手法では決まった品詞の連続パターン (以下、複合語パターンと略記) を複合語であるとみなし、抽出する。図 4.4 は、名詞の連続を複合語とみなした場合の例である。まず、複合語は 2 単語以上で構成されなくてはならない。この例では、単語「登場」が名詞であるため複合語パターンと合致するが、次に続く単語「する」が動詞のため、抽出対象から除外される。

探索は「する」の次の単語である、「人」より再開される。再開地点からは「人」、「型」、「ロボット」と名詞が続いており、その後助詞の「の」が検出された時点で、「人型ロボット」が複合語として抽出される。

複合語パターンだけではヒント文に用いる複合語としてふさわしくないものも抽出される。そこで本手法では複合語パターンの他、以下の条件にあてはまるものを複合語抽出の対象から除外した。

- 数・地域-一般・人名
- アルファベットを含むもの
- 平仮名のみで構成されるもの
- 一度しか出現しないもの

抽出した複合語は出現回数順にソートし、データベースに保存する。後からキーワードにヒント文を付与する際にはまず 1) データベースからキーワードを含む複合語の内出現回数の最も多い複合語を検索し、2) 検索された複合語を含む 1 文をランダムに抽出する。

ここでは出現回数の最も多い複合語しか用いないが、これを出現回数の多い順に複数個を用いることで、抽出されるヒント文に変化を持たせられることを付記しておく。

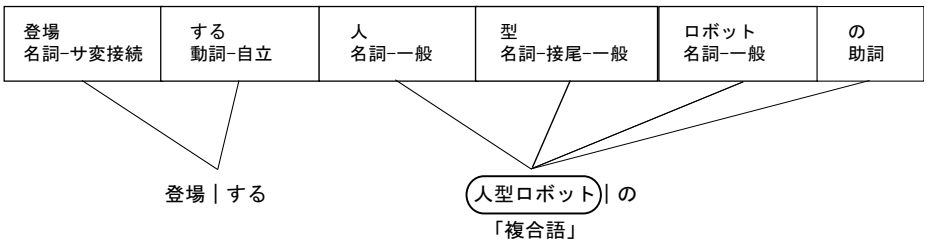


図 4.4: 名詞の連続を抽出する例

4.3 連想形

4.3.1 単語による連想形

単語による連想形を以下に示す。

C.　ナイル、アマゾン、信濃　川 (カワ)

キーワードと関連の深い単語を複数列挙するヒント文を、単語による連想形と呼ぶ。上記の例ではキーワード「川」と関連の深い「ナイル」「アマゾン」「信濃」の 3 単語がキーワードの手掛かりとなるイメージを作りだしているのがわかる。

単語間の関連の深さを測る尺度として、本手法では相互情報量を導入する。相互情報量の式は以下のようになる。

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{\frac{DF(x,y)}{N}}{\frac{DF(x)}{N}\frac{DF(y)}{N}} = \log \frac{N \cdot DF(x,y)}{DF(x)DF(y)}$$

ここで、$P(x,y)$ は全体の文集合において単語 x,y を共に含む文書の割合、$P(x)$ は単語 x を含む文の割合、$P(y)$ は単語 y を含む文の割合である。また、N は全体における文数、$DF(x,y)$ は x,y を共に含む文数、$DF(x), DF(y)$ はそれぞれ x,y を含む文数である。

単語による連想形を生成する前段階として、まず 1) 上式を用いてあらゆる 2 単語間の相互情報量を計算し、続いて 2) 相互情報量の大きい順にソートし、データベースに保存しておく。後からキーワードにヒント文を付与する際には、データベースからキーワードとの相互情報量が高い上位 3 単語を検索する。

4.3.2 文による連想

文による連想形を以下に示す。

C. 日の当たる場所で○○○○○○。ポカポカしていい気持ち。　　日向ぼっこ (ヒナタボッコ)

文による連想形とは、キーワードと関連の深い単語を文中に埋め込むことで、キーワードを推測するための文脈を形成するヒント文である。C の例では、「ポカポカしていい気持ち。」部がキーワードと関連の深い文である。「ポカポカ」・「いい気持ち」というキーワードと関連の深い単語を文中に埋め込むことで、キーワードを推測させている。関連文による連想形を生成する手順としては、まず 1) 関連の深い単語を選び、2) 関連の深い単語を含む文を抽出する。このうち前者の工程は、(4.4) 節で説明した手法と同じである。

ここでは後者の、関連の深い文の計算方法を説明する。関連の深さを表す式を (4.2)(4.3) に示す。

$$sh(i,k) = \frac{1}{|h_k|} \sum_{w_j \in h_k} I^+(w_i, w_j) \tag{4.1}$$

$$I^+(x,y) = \begin{cases} I(x,y) & I(x,y) \geq 0 \text{ の場合} \\ 0 & I(x,y) < 0 \text{ の場合} \end{cases} \tag{4.2}$$

ここで w_i はヒント文を付与されるキーワード、h_k は k 番目のヒント文候補を表す。(4.4) 節で計算した単語間の相互情報量を用い、あるヒント文候補 h_k の w_i との関連の深さ $sh(i,k)$ を、h_k 中に存在する w_i 以外の全ての単語と w_i との相互情報量の和と定義する (ただし相互情

報量が負になる組は 0 とする)。ただし $|h_k|$ はヒント文候補 h_k に含まれる単語数である。ここで計算した $sh(i,k)$ の値が最も高い 1 文を、ヒント文として与えられたキーワード w_i に付与する。

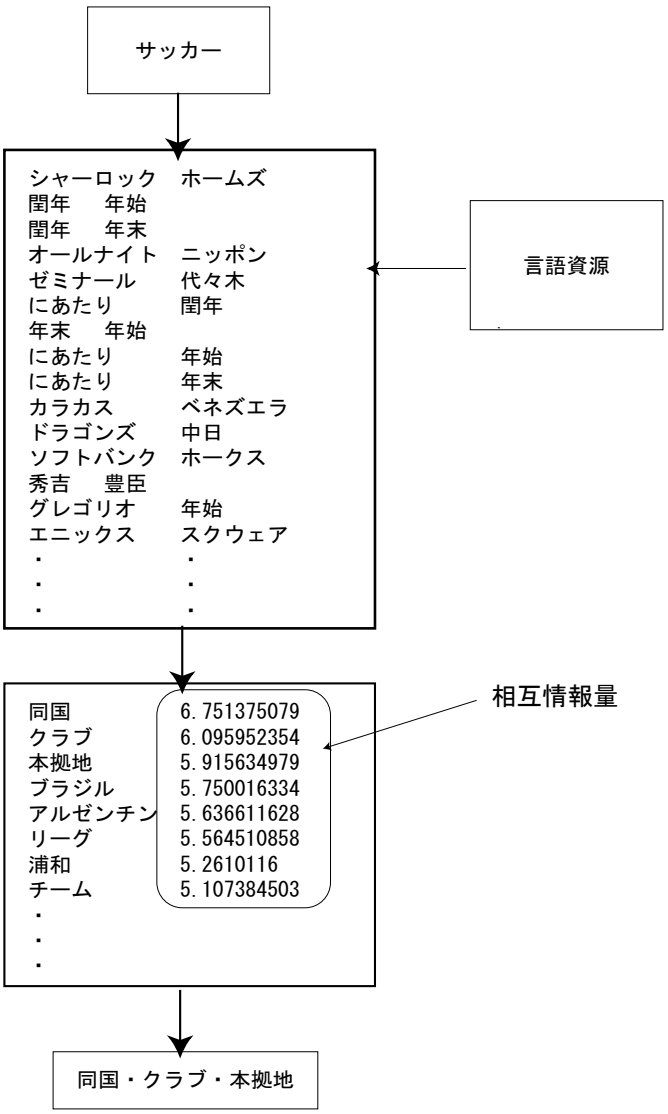


図 4.5: 単語による連想形ヒント文抽出の流れ

第5章　実験と考察

5.1　実験の概要

表 5.1: 実験データの詳細

	学研 現代新国語辞典	ウィキニュース	Wikipedia Abstract	毎日新聞	ブログ
記事数	-	1787	709101	98,211	127,456
単語数	65,000[1]	-	126,150	132,772	1,547,914

本章では、4 章で述べた生成手法を実験し、生成されたヒント文を評価・考察する。本研究では (3.1) 節で作成したクロスワードヒントデータベースから 100 個のキーワードをランダムに抜きとり、それぞれの手法が 100 個中いくつのキーワードにヒント文を付与できるかを評価する。その後、生成されたヒント文のうちいくつがクロスワードヒントとして使えるかを人手で評価し、成功例、失敗例について原因を考察する。以下ではクロスワードヒントデータベースからランダムに抽出したキーワード群を評価キーワードと呼ぶ。

ヒント文生成に用いたコーパスは、「学研 現代新国語辞典」、「ウィキニュースの記事タイトルデータベース[2]」、「Wikipedia Abstract[3]」、「毎日新聞 1999 年 CD-ROM 版」、「Excite ブログ」の 5 つである。

使用するコーパスデータの詳細を、表 5.1 にまとめる。単語分割には茶筌を用いた。

5.2　説明文の生成

5.2.1　辞書を用いた説明文の生成

説明文の生成には辞書を用いる。本研究では説明文生成の実験データとして、「学研 現代新国語辞典 金田一春彦編 初版 (c)1994」(以下、辞書と略記) を用いた。収録語数は 65000 語である。辞書の見出語は漢字で表記されていなかったため、例えば「奥 (オク)」というキーワードに対して「億」「屋」「置く」に関する説明文も付与されることを付記しておく。

生成されたヒント文の例を表 5.2 に示す。評価キーワード 100 個を対象にヒント文の付与を実験したところ、69 個にヒント文が付与され、生成されたヒント文の総数は 437 文であっ

[2]http://download.wikimedia.org/jawikinews/20061222/

[3]http://download.wikimedia.org/jawiki/20061220/jawiki-20061220-abstract.xml

表 5.2: 辞書を用いた説明文生成の成功例

ヒント文	キーワード
けがれを払うため、湯や水で体を清めること。	行水 (ギョウズイ)
その物を光がよく通り、向こうがすきとおってみえること。	透明 (トウメイ)
少しの間降ってすぐに晴れる雨。	通り雨 (トオリアメ)
動力によって人や荷物を上下に移動させる機械。	エレベーター
記号は、G。	ガウス

表 5.3: 辞書を用いた説明文生成の失敗例

番号	失敗例	キーワード
1	軽い驚きを表す語。	運 (ウン)
2	あらかじめ…する。	奥 (オク)
3	肉食性。	スカンク (スカンク)

た。また、付与されたヒント文が実際にクロスワードで使えるかを人手で評価した結果、437 文中 249 文が有用なヒント文であった。

ヒント文として使えないものに多かったのが、キーワードの同音異義語を説明するものであった。今回の評価では、同音異義語を説明していてもそれがクロスワードヒントとして使える内容であればよしとしたが、同音異義語の説明文が付与される可能性は別の問題も孕んでいる。例えば 1 の失敗例で与えられたキーワードは名詞の「運 (ウン)」であるが、付与されたヒント文は感動詞の「うん (ウン)」を説明してしまっている。キーワードに用いる単語は名詞が原則であるため、たとえヒント文「軽い驚きを表す語。」が感動詞の「うん (ウン)」の説明文として成立したとしても、このヒント文はクロスワードに使うには不適切である。これはキーワードを与える際に同音異義語を区別できることができれば、回避可能な問題である。同様のことが、2 の失敗例にもいえる。こちらではキーワードとして名詞の「奥」を与えられているが、ヒント文は動詞の「おく」を説明してしまっている。

次に、3 の失敗例を考察する。このヒント文はキーワードを推測する手掛かりとしてあまりに弱いため、ヒント文として使えないと評価した。肉食性の動物はスカンクの他にも山ほど存在するうえ、「肉食性」という性質はスカンクを説明する上で重要な性質ではない。

5.3 虫食い文の生成

5.3.1 辞書を用いた虫食い文生成

「学研　現代新国語辞典」を用いて、虫食い文を生成した。文切りは句点区切りで行い、単語分割には茶筌を用いた。生成されたヒント文の例を表 5.4 に示す。(4.1) 節の手法を実験した結果、評価キーワード 100 個中、47 個にヒント文が付与された。辞書は 1 つのキーワードに対し、複数のヒント文を付与する場合がある。生成された虫食い文の総数は 213 文で、そ

表 5.4: 辞書を用いた虫食い文生成の成功例

ヒント文	キーワード
「烏の○○○○○」	行水 (ギョウズイ)
「○○○○不明で返送された」	宛先 (アテサキ)
「犬も歩けば○○に当たる」	犬 (イヌ)

表 5.5: 辞書を用いた虫食い文生成の失敗例

番号	失敗例	キーワード
1	「お○○○する」	邪魔 (ジャマ)
2	「○○、めずらしい人がきた」	親 (オヤ)
3	「念頭に○○」	奥 (オク)

の内の 134 が文ふさわしいヒント文であった。

ヒント文として不適切と評価した例を表 5.5 に示す。失敗例として最も多かったのが、キーワードを推測するに足るだけの情報がないケースである。1 の失敗例では、「邪魔」というキーワードに「お○○○する」というヒント文が付与された。しかし、「お○○○する」のように表現される単語は、例えば「お休みする」「お願いする」等を筆頭に数多く存在するため、この情報から「邪魔」を推測するのは甚だ困難である。このような記述では、辞書は見出語が用いられる形式を例示しているにすぎず、見出語に特有の表現を述べているわけではない。よって 1 の失敗例のように、キーワードを絞りこむのが困難な虫食い文が多く生成させた。

説明文の生成の場合と同様に、与えられたキーワードの同音異義語の説明文が付与される場合も多くみられた。2 の失敗例では名詞で与えられた「親」が感動詞の「おや」、3 の失敗例では名詞で与えられた「奥」が動詞の「置く」として虫食い文が付与されている。

5.3.2 ニュース記事のタイトルを用いた虫食い文生成

ニュース記事のタイトル内の単語どうしはお互いに密接な関りを持つため、タイトル中の単語の 1 つを虫食いにした場合に他の単語から得られるキーワードを推測するための情報が豊富である。本手法では、虫食い文生成の実験データに「ウィキニュース　12 月 22 日付け記

表 5.6: ニュース記事のタイトルを用いた虫食い文生成の成功例

ヒント文	キーワード
ビルゲイツ氏、2008 年 7 月に○○○○○○○から退く	マイクロソフト (マイクロソフト)
○○○○イーグルス・田尾監督を解任	楽天 (ラクテン)
ガソリン○○○が値上げ - 140 円台に	価格 (カカク)

表 5.7: ニュース記事のタイトルを用いた虫食い文生成の失敗例

番号	失敗例	キーワード
1	愛知県で○○○が大量に盗まれる	タバコ (タバコ)
2	甲子園に大鉄○○復活へ	傘 (カサ)
3	○○○ワの自爆テロ計画、男性拘束	サマー (サマー)

事タイトルデータ[4]」を用いた。記事数は 1762 記事である。文切りは句点を基準にし、単語分割には茶筌を用いた。(4.2) 節で説明した手法を実験した結果、評価キーワード 100 個中、23 個にヒント文が付与された。1 つのキーワードが複数の記事に存在することがあるため、1 つのキーワードに対し、複数のヒント文が付与される場合がある。生成されたヒント文の総数は 73 文であった。

虫食い文生成の成功例を表 5.6 に示す。3 つの例はいずれも、キーワードと関連の深い単語が文中に埋め込まれているか、キーワードが文中で複合語を構成する単語の 1 つとして使われていることで、文中の他の部分からキーワードを推測するに足る量の情報を得ることができる。虫食い文生成の失敗例を表 5.7 に示す。ニュース記事のタイトル特有の失敗例として、問題が出された時点の状況によってヒント文としての適切さが変化するものがある。1 の失敗例は愛知県でタバコが盗まれた事件が話題にのぼらない時点で評価したために失敗例の 1 つとしたが、例えば愛知県でタバコが盗まれた事件が起き、ニュースを賑わしている時点でこのヒント文を評価すると、このヒント文は成功例に分類することができる。たとえ文中にキーワードと普遍的に関連の深い単語が存在しなくとも、ヒント文が提示された時点で関連の深い単語が文中に埋め込まれているならば、それは適切なヒント文となりうる。本実験ではこのようなヒント文を「どちらともいえない」と分類した。ニュース記事のタイトルには固有名詞が頻出するため、茶筌による単語切り出しや単語の読み間違いが目立った。2 の失敗例では、正しくは「大鉄傘 (ダイテッサン)」と読むべき単語を茶筌が「ダイテツカサ」と読んでしまったために、キーワードの読みと文中におけるキーワードの読みが違ってしまっている。また、3 の失敗例では地名である「サマーワ」を「サマー」と「ワ」に単語分割してしたために、キーワードが文中で、単語として扱われていない。本実験ではこのようなヒント文を「不適切である」と分類した。ウィキニュースによって生成したヒント文を人手で評価したところ、適切なものが 51 文、不適切なものが 15 文、どちらともいえないものが 7 文という結果となった。

[4]http://download.wikimedia.org/jawikinews/20061222/

5.3.3 複合語を用いた虫食い文生成

実験データ

本手法において実験に用いたコーパスは、「Wikipedia Abstract 12 月 20 日付けデータ[5]」、「毎日新聞 1999 年 CD-ROM 版」、「Excite ブログ[6]」の 3 つである。3 種のコーパスに対して実験を行った理由は、それぞれのコーパスの特性によって生成されるヒント文にどのような違いが生まれるかを知るためである。Excite ブログのコーパス作成にはクローラとして wget を使用し、2005 年 8 月 4 日から約 1 週間、1 つのシードページから 20,275 サイト　690,184 記事をクロールし、それぞれの記事から body 部を抽出、さらに本文中の html タグを全て取り払ったものをコーパスとして用いた。使用するデータ内容を表にまとめる。文切りは句点を基準とし、単語分割には茶筌を用いた。

予備実験

表 5.8: 複合語抽出結果

品詞パターン	抽出した複合語数	適切なものの割合
名詞+未知語	3489 語	20/100
名詞のみ	37205 語	76/100
名詞のみ・名詞+未知語の連続	39845 語	70/100

表 5.9: 未知語を含んだ複合語の抽出例

番号	品詞パターン	複合語
1	NU	名犬 ラッシー
2	UN	ジオン 抗争
3	UN	カピス 州
4	UN	アア 溶岩

(4.3) 節で述べた手法は、まず 1) 複合語を抽出し、続いて 2) 複合語を含む文を抽出する。ここでは予備実験として、3 つの複合語抽出方法の比較・検討を行う。以下に示す 3 つの品詞パターンの連続を、複合語として検出する。

- 名詞+未知語
- 名詞のみ

[5]http://download.wikimedia.org/jawiki/20061220/jawiki-20061220-abstract.xml
[6]http:www.exblog.jp/

- 名詞+未知語・名詞のみの両方

実験データには「Wikipedia Abstract」の 12 月 20 日付けデータを用いた。

複合語抽出の結果を表 5.8 に示す。未知語を混ぜることで複合語の抽出数は増えたが、それに伴って適切な複合語の割合は減少した。「名詞+未知語」パターンにより抽出したものを表 5.9 に示す。U は未知語、N は名詞を表す。1・2 のアニメ用語等、サブカルチャー的な用語を切り出せるのが、未知語を含めた場合の魅力であるが、一方で 3・4 のように、確かに存在はするが、一般には誰も知らない地名や用語が抜き出されることも多い。

抽出する複合語はキーワードを推測させる手掛かりを持たなければならない。、よって抽出する複合語は一定以上の認知度がない限り、適切でないと評価する。本研究では、適切なヒント文をなるたけ多く生成することを重視する。したがって複合語による虫食い文を生成する際には、「名詞のみ」の連続によって複合語を切り出すことにする。

実験結果と考察

評価キーワード 1 個につき、ヒント文を 1 文付与した。なお、抽出の際に括弧で囲まれた記述はすべて削除した。Wikipedia Abstract・毎日新聞・excite ブログで行った虫食い文生成の評価を、表 5.10 に、生成したヒント文の成功例をそれぞれそれぞれ表 5.11・表 5.12・表 5.13 に示す。

ヒント文として不適切なものは、例えば「お受験では母親が子供を塾に通わせますが、そ

表 5.10: 実験結果

コーパス	生成ヒント数/100	適切	どちらともいえない	不適切
Wikipedia Abstract	57/100	34/57	7/57	14/57
毎日新聞	67/100	49/67	8/67	9/67
Excite ブログ	77/100	49/77	13/77	15/77

表 5.11: Wikipedia Abstract による虫食い文の成功例

番号	ヒント文	キーワード
1	学術用語としては、「蛋白質」という○○○表記は用いず、「タンパク質」と表記する	漢字 (カンジ)
2	テンプシーロールとは、ボクシングの元世界ヘビー級王者ジャック・テンプシーが編み出した必殺○○○	ブロー (ブロー)
3	○○○問題とは、鯨およびイルカの捕獲の是非に関する論争、国内外の摩擦問題である。	捕鯨 (ホゲイ)

表 5.12: 毎日新聞による虫食い文の成功例

番号	ヒント文	キーワード
1	○○勘定なら、今の千代大海には 14 勝が必要だ。	星 (ホシ)
2	長野五輪での大活躍も、連盟の○○○○具合には直結しなかったということか。	懐 (フトコロ)
3	京都直送の「生○○刺身」などさっぱりしたメニューも用意した。	湯葉 (ユバ)

表 5.13: Excite ブログによる虫食い文の成功例

番号	ヒント文	キーワード
1	読み間違ってるままひらがな入力し、○○○変換したんやろな…	漢字 (カンジ)
2	恐怖映画の○○○的傑作「ゾンビ」を、CM 出身の新鋭ザック・スナイダーが監督した	歴史 (レキシ)
3	現在発売中の○○○ジャンボ宝くじ!!	サマー (サマー)

の送り迎え、家庭での予習○○○○○、生活全般において、コーチのようにベッタリと子供に付いていないとうまくいかないところもあるんです。(キーワード:復習 (フクシュウ)」のように、明らかに冗長なものが多くみられた。本手法では句点毎に区切ったものを機械的に抽出しており、文の長さを考慮しなかった。この問題は、例えば「文中の読点は 2 回以上許さない。」「9 単語以上の文はヒント文候補から除外する。」といったルールを新たに導入することで、ある程度の解決ははかれると考える。

その他の原因による失敗例には、「-関連○○-(キーワード:記事 (キジ))」のように文になっていないものや、「アミロースとは、多数の α -グルコース分子がグリコシド結合によって重合し、直○○○状になった高分子である (キーワード:鎖 (クサリ))[7]」のように、茶筌に読み間違いによるものがちらほらみられた。

5.4 連想形

5.4.1 辞書を用いた単語の羅列による連想形

表 5.14: 辞書を用いた単語の羅列による連想形の成功例

ヒント文	キーワード
入り日・夕日・落日	洛陽 (ラクヨウ)
川岸・州・岸	瓦 (カワラ)
チェーン・つながり・きずな	鎖 (クサリ)

「学研 現代新国語辞典」を用いて、単語の羅列による連想形を生成した。(4.1) 節で述べた手法を実験し、単語として出力されたものの中からランダム 3 単語を選出し、スリーヒント[8]形式のヒント文を生成した。この際、辞書から単語が 3 つ以上抜き出せなかった場合はヒント文が付与されなかったとした。評価キーワードへのヒントの付与を実験した結果、100 個中 22 個にヒント文が付与された。このうち 8 文が、ヒント文として有用なものであった。評価の際、与えられたキーワードの意味は考えず、読みが推測できればよいものとした。生成されたヒント文の例を表 5.14 に示す。

辞書データに漢字の見出しがないため、抽出された単語はそれぞれが異なる意味の単語を説明していることが度々あった。例えば「値段・来客・旅客 (キーワード:価格 (カカク))」の

[7]直鎖状 (チョクサジョウ)
[8]キーワードと関連の深い単語を 3 つ列挙する形式のヒント文。

スリーヒントでは、「値段」は「価格」を説明しているが、「来客」・「旅客」は「過客」を説明している。このように異なる意味を説明する単語が混在したために、キーワードを推測するのが困難なものが多かった。

5.4.2 相互情報量を用いた単語の羅列による連想形

表 5.15: 実験結果

コーパス	生成ヒント文数/100	成功文数
Wikipedia Abstract	26/100	11/28
毎日新聞	50/100	22/50
Excite ブログ	56/100	27/56

表 5.16: Wikipedia Abstract を用いた単語の羅列による連想形の成功例

番号	ヒント文	キーワード
1	オペレーティングシステム・ページ・ソフトウェア	マイクロソフト (マイクロソフト)
2	遺伝・配列・染色	遺伝子 (イデンシ)
3	ボール・エニックス・題名	ドラゴン (ドラゴン)

表 5.17: 毎日新聞を用いた単語の羅列による連想形の成功例

番号	ヒント文	キーワード
1	量販・エアコン・冷蔵庫	家電 (カデン)
2	ウィンドウズ・OS・ソフトウエア	マイクロソフト (マイクロソフト)
3	マリナーズ・オリックス・外野	イチロー (イチロー)

表 5.18: Excite ブログを用いた単語の羅列による連想形の成功例

番号	ヒント文	キーワード
1	生態・外資・交感神経	系 8 (ケイ)
2	イーグル・岩隈・一場	楽天 (ラクテン)
3	アレルギー・西日本・太り	体質 (タイシツ)

「Wikipedia Abstract 12 月 20 日付けデータ」、「毎日新聞　1999 年 CD-ROM 版」、「Excite ブログ」の 3 種のコーパスを用いて、(4.4) 節で説明した手法を実験した。コーパスの種類毎に、実験の評価結果を表 5.15 に、ヒント文の成功例をそれぞれ表 5.16、5.17、5.18 に示す。

コーパス内に表れる単語全てを対象に相互情報量を計算したため、例えば「ラフ」のスリーヒントとして「ァ・エル・ひどかっ」の 3 つが出力された。これは単語「ラファエル」を茶筌

が「ラフ」「ァ」「エル」と切り出したためと思われる。この他「煙草」のスリーヒントに「吸い・吸わ・吸っ」と、原形が同じものが活用が違うだけでスリーヒントとして出力されることもあった。

前者は、相互情報量を計算する対象を名詞と形容詞だけにするなどすることで対処可能である。後者は、原形が同じものを 1 つにまとめることで、解消される。

第6章　おわりに

本稿では、専門雑誌で掲載されている日本語クロスワードを収集・分類することでクロスワードヒントの性質を把握し、それに基づいて辞書やオンラインニュース、各種コーパス、ブログからヒント文を抽出する手法を提案した。

実際の日本語クロスワードの性質を把握する目的で、専門雑誌に掲載されていたクロスワードヒント 984 文をランダムに収集し、分析・分類を行った。その結果、日本語クロスワードで用いられるヒント文はほぼ 1 文で構成され、その分類タイプは説明文・虫食い文・連想形に属するものが高い割合を占めることがわかった。そこで、本研究ではこれらの性質を持つヒント文を生成するため、1) 辞書から説明文・単語・虫食い文を抽出する手法、2) ニュースの見出しから虫食い文を生成する手法、3) 各種言語資源から複合語による虫食い文を生成する手法、4) 相互情報量を用いて、単語による連想形ヒント文を生成する手法 5) 相互情報量を用いて、文による連想形ヒント文を生成する手法を提案し、そのうち 1)〜4) の手法を実験・評価した。その結果、自動的にヒント文を生成するまでにはいたらなかったが、提案手法によって人間の行うクロスワードの作成を支援できる可能性を示した。

今回提案した手法はまだ発展段階にあり、新たな工夫を加えることによってより精度のよいヒント文生成が可能となる。また今後の課題として、本稿では実験を行わなかった 5) の手法についても、実験・考察する必要がある。

謝辞

本研究を進めるにあたり、多大なる御指導をいただきました筑波大学システム情報工学研究科山本幹雄助教授に心から感謝いたします。また、貴重な御意見、御協力をいただきましたシステム情報工学研究家椎名毅教授、同滝沢穂高講師、同山川誠講師に厚く御礼申し上げます。最後に、知能情報研究室の皆様から様々な手助けと御協力をいただきました。心から感謝いたします。

付 録A　評価キーワード

表 A.1: 評価キーワード一覧

番号	キーワード (読み)	キーワード
001	ギョウズイ	行水
002	ラクヨウ	洛陽
003	タイ	タイ
004	ケイ	系
005	クツシタ	靴下
006	カンジ	漢字
007	カワラ	瓦
008	レキシ	歴史
009	ミノ	みの
010	ヤクシ	訳詞
011	ウショウ	鵜匠
012	ブロー	ブロー
013	セイコ	聖子
014	クッタク	屈託
015	アジア	アジア
016	クサリ	鎖
017	ユカリ	ゆかり
018	チュウカ	中華
019	イドウケイサツ	移動警察
020	ラフ	ラフ
021	ジャマ	邪魔
022	キアイ	気合い
023	トウメイ	東名
024	ホシ	星
025	アテサキ	宛先
026	ウン	運
027	コエ	声
028	オヤ	親
029	ヤオヤ	八百屋
030	スカンク	すかんく
031	ホヨウ	保養
032	カデン	家電
033	ウエット	ウエット
034	ダカイ	打開
035	ホゲイ	捕鯨
036	オク	奥
037	イロメガネ	色眼鏡
038	フトコロ	懐
039	マイクロソフト	マイクロソフト
040	ラクテン	楽天
041	タイシツ	体質
042	シツウ	歯痛
043	ウシミツ	丑三つ
044	アイルランド	アイルランド
045	キジ	記事
046	トリック	トリック
047	トオリアメ	通り雨
048	イデンシ	遺伝子
049	ボサツ	ぼさつ
050	ナンイド	難易度

番号	キーワード (読み)	キーワード
051	フクシュウ	復習
052	アタマ	頭
053	エレベーター	エレベーター
054	バイショウ	賠償
055	チユ	治癒
056	フウリン	風鈴
057	イキガイ	生きがい
058	クジ	くじ
059	タバコ	煙草
060	ガウス	ガウス
061	ツラ	面
062	キバ	牙
063	タク	卓
064	マージャン	麻雀
065	ドラゴン	ドラゴン
066	ブシ	武士
067	チノウハン	知能犯
068	ショメイ	署名
069	ユバ	湯葉
070	アカ	紅
071	スラング	スラング
072	ケショウ	化粧
073	イチロー	イチロー
074	サンポ	散歩
075	ポルトガル	ポルトガル
076	ウオ	魚
077	メンルイ	麺類
078	アルカイダ	アルカイダ
079	キー	キー
080	カサ	傘
081	カミノケ	髪の毛
082	ソウカンズ	相関図
083	カエリミチ	帰り道
084	ヤミヨ	闇夜
085	スマッシュ	スマッシュ
086	コカゲ	木陰
087	ガン	雁
088	キャタツ	脚立
089	スウコウ	崇高
090	カカク	価格
091	マウス	マウス
092	キナコ	きなこ
093	フクシ	福祉
094	スウジ	数字
095	キリツ	起立
096	シソ	紫蘇
097	サマー	サマー
098	ボウ	棒
099	クライ	位
100	デンワチョウ	電話帳

参考文献

[1] 遠山顕. 遠山顕のクロスワードの謎. 日本放送出版協会,2002.11

[2] Keim,G.A.,Shazeer,N.M.,Littman,M.L,Agarwal,S.,Cheves,C.M.,Fitzgerald,J.,Grosland,J.Jiang,F.,Pollard,S. and Weinmeister,K. PROVERB;The Probabilistic Cruciverbalist *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp.710-717(1999).

[3] Shazeer,N.M.,Littman,M.L. and Keim,G.A. Solving Crossword Puzzles as Probabilistic Constraint Satisfaction *Proceedings of the Sixteenth National Conference on Artificial Intelligence*,pp.156-162(1999).

[4] Litttman,M.L.,Keim,G.A. and Shazeer,N.M. Solving Crossword with PROVERB *Proceedings of the Sixteenth National Conference on Artificial Intelligence*,pp.914-915(1999).

[5] Berghel,H.,Yi,C. Crossword compiler compilation. *The Computer Journal 30*, pp.276-280, 1989.

[6] Aoife Aherne and Carl Vogel. Crossing WordNet with Crosswords,Netting Enhanced Automatic Crossword Generation. *Trinity College technical report*, 05-July-2005.

[7] Keim,G.A.,Shazeer,N.M.,Littman,M.L.,Agarwal,S.,Cheves,C.M.,Fitzgerald,J.,Grosland,J.,Jiang,F.,Pollard,S. and Weinmeister,K. PROVERB:The Probablistic Cruciverbalist. *Proceedings of the Sixteenth National Conference on Artificial Intelligence*,pp.710-717(1999).

[8] 加部通明, 生方俊典. クロスワードパズルの遺伝的アルゴリズムによる作成. 第 52 回平成 8 年前期情報処理学会全国大会講演論文集,No.2,pp.133-134,1996.

[9] 帆苅譲, 石川勉, 笠原要. 概念ベースを用いたクロスワードパズル作成システム. 第 56 回平成 10 年前期情報処理学会全国大会講演論文集,No.2,pp.312-313,1998.

[10] 藤井敦, 伊藤克亘, 秋葉友良. 事典的 Web 検索サイトの構築. 言語処理学会第 9 回年次大会発表論文集,pp.129-132,2003.

[11] 芳鐘冬樹、野澤孝之、辻慶太、影浦狭. ウェブからの関連語・下位語の収集手法の検討と検索システムへの応用第 52 回日本図書館情報学会研究大会発表要綱,pp.113-116,2004.

[12] 佐藤理史. 日本語クロスワードパズルを解く. 情報処理学会研究報告,Vol.2002,No.4,2002-NL-147-. 11,pp69-76,2002.