1U-9

日本語クロスワードパズルを解く：辞書逆引きにおける検索語の重みの決定法

内木 賢吾 † 佐藤 理史 † 駒谷 和範 †

† 名古屋大学大学院 工学研究科 電子情報システム専攻

1. はじめに

クロスワードパズル (クロスワード) は，世界中でよく知られた言語パズルのひとつである．クロスワードを解くためには，カギとよばれるヒントから答となるワードを推定し，グリッドとよばれるマス目を正しく埋める必要がある．我々は，日本語クロスワードを人間並みに解くシステムの実現を目指し，2011 年度よりシステムの開発を行っている [1][2].

カギは，どのような方法で答 (ワード) を推測できるかによって，いくつかの種類に分類することができる [2]. そのひとつに，辞書の逆引きによって答が推測できるカギがある．ここで，辞書の逆引きとは，カギとよく似た (類似度が高い) 定義文の見出し語を探すことである．例えば，「魚の体を覆っている小薄片」というカギに対し，それと類似する辞書の定義文「魚類・爬虫類などの体の表面をおおっている小片」を見つけることができれば，「ウロコ」という正しい答を導くことができる．

辞書の逆引きに用いる類似度は，カギと定義文の両方に出現する語とそれらの語の重みに基づいて計算する方法が一般的である．この重みは通常，語の出現頻度に基づいて計算される (出現頻度の少ない語を重要視する) [3][4] が，この方法では，語の重要度はそれぞれの語に対して一定となる．しかしながら，実際のカギを観察すると，語の重要度をそれぞれのカギに応じて決定したほうがよいと思われる場合がある．

本論文では，与えられたカギから，そこに含まれる語の重要度を，語の出現順序と後続助詞から決定する方法を提案し，既存の方法 (TF-IDF 法) より優れていることを実験的に示す．

2. 検索語の重み決定法

2.1 辞書の逆引き

辞書の逆引きは，与えられたカギに対して，スコア付き候補リストを出力する．検索対象辞書には，岩波国語辞典 第 5 版 (57,618 語)，EDR 日本語単語辞書 (194,019 語)，日本語 Wikipedia (803,229 語) を用いる．辞書の逆引きは，次の 3 ステップから構成される．

(1) 検索語の抽出
カギ c を MeCab+Unidic で形態素解析し，内容語の語彙素 $q \in Q(c)$ をすべて抽出する．ここで，語彙素とは，Unidic で定義された階層的見出しの最上位の階層に位置する語である．得られた語彙素を，以下では検索語と呼ぶこととする．さらに，文字列 c から得られる検索語リストを $Q(c)$ と書く．

(2) (各辞書に対する) スコア付き候補リスト生成
辞書 D から，検索語 $q \in Q(c)$ を含む定義文 d をすべて検索し，それらの見出し語を答の候補とする．

表 1: 重要な語の調査

出現文節	カギ数		後続助詞	カギ数	
1	31	70%	格助詞 (+動詞)	20	45%
2	4	9%	格助詞	17	39%
3	4	9%	係助詞	4	9%
4	3	7%	並列助詞	2	5%
5 以上	2	5%	副助詞	1	2%
計	44	100%	計	44	100%

各候補のスコアは，以下の式で計算する．

$$\mathrm{score}(D,c,d) = \alpha \cdot \frac{\sum_{r \in Q(c) \cap Q(d), d \in D} w(c,r)}{\sum_{q \in Q(c)} w(c,q)} \quad (1)$$

ここで，α は候補リストのスコアの総和を 1 とするための正規化係数，$r \in Q(c) \cap Q(d)$ はカギ c と定義文 $d \in D$ に共通して含まれる検索語，$w(c,r)$ は検索語 r のカギ c における重みを表す．このスコアは，重みが大きい語が多く一致するほど高くなる．

(3) 複数の候補リストのマージ
それぞれの辞書 $D \in \mathscr{D}$ に対して得られた複数の候補リストを，1 個の候補リストに集約する．このとき，候補 t の最終スコア $\overline{\mathrm{score}}(\mathscr{D},c,t)$ を以下の式で計算する．

$$\overline{\mathrm{score}}(\mathscr{D},c,t) = \overline{\alpha} N(t)^2 \max_{D \in \mathscr{D}} (\beta(D) \cdot \mathrm{score}(D,c,t)) \quad (2)$$

ここで，$\overline{\alpha}$ は集約された候補リストのスコアの総和を 1 とするための正規化係数，$N(t)$ は候補 t を含む候補リストの数，$\beta(D)$ は辞書 D の優先度を表す係数である．β の値は，現在は，岩波国語辞典 ($\beta = 6$)，EDR 日本語単語辞書 ($\beta = 3$)，日本語 Wikipedia($\beta = 1$) を用いている．

2.2 $w(c,q)$ の決定法

検索語の重み $w(c,q)$ の計算法の決定に先立ち，実際のクロスワードのカギにおいて，どのような語が重要と考えられるかについて調査した．この調査には，2009 年と 2011 年の『JAF Mate』に掲載されたクロスワード 20 問に含まれるカギ計 512 個中，44 個を使用した．調査結果を表 1 に示す．この表において，出現文節は，重要と思われる内容語がカギの先頭から何文節目に出現するかを示す．後続助詞は，重要と思われる内容語に後続する助詞の品詞細分類を表す．この表より，内容語がカギの先頭に近い文節に出現する場合，および，後続助詞として格助詞を取る場合に，重要度が高い語となりやすいことがわかる．

Solving Japanese Crossword Puzzles: Determining Weights of Query Words for Reverse Dictionary Lookup, *Kengo Naiki (Nagoya Univ.), Satoshi Sato (Nagoya Univ.), and Kazunori Komatani (Nagoya Univ.)*

表 2: 実験結果

データセット	重み	カギ数	正解を含む	Recall(*N*) 1	5	10	50	MRR	平均スコア	ワード正解率
朝日新聞	baseline	1,003	79.6%	21.2%	39.2%	47.4%	65.1%	0.380	0.0883	-
	$w_o(c,q)$			28.6%	48.2%	55.0%	67.1%	0.472	0.0892	(87.8%)
	$w_p(c,q)$			28.8%	48.0%	54.8%	67.4%	0.475	0.0895	(87.8%)
	$w_c(c,q)$			28.7%	47.5%	54.5%	66.9%	0.470	0.0911	(89.0%)
JAF Mate	baseline	910	60.2%	12.7%	25.4%	33.0%	46.6%	0.314	0.0738	-
	$w_o(c,q)$			16.9%	29.5%	34.3%	47.1%	0.383	0.0827	(83.3%)
	$w_p(c,q)$			17.6%	29.0%	34.6%	48.2%	0.390	0.0802	(83.2%)
	$w_c(c,q)$			17.1%	28.9%	34.0%	46.4%	0.380	0.0854	(83.6%)

カギ c	:	魚	の	体	を	覆っ	て	いる	小	薄片
検索語 q	:	魚		体		覆う		居る		薄片
(a) $w_o(c, q)$	:	4		3		2		2		1
(b) $w_p(c, q)$	:	3		3		3		0.5		1
(c) $w_c(c, q)$	:	12		9		6		1		1

図 1: 重みの実例

この結果に基づき，我々は，カギにおける出現順序と後続助詞の種類という 2 つの観点を考慮した，以下の 3 種類の検索語の重み決定法を提案する．これらの方法の具体例を図 1 に示す．

(a) 出現文節に基づく重み付け w_o
検索語 q が文末から n 文節目に含まれる場合

$$w_o(c,q) = n \tag{3}$$

(b) 品詞，後続助詞に基づく重み付け w_p

$$w_p(c,q) = k \tag{4}$$

- 検索語 q の品詞が，固有名詞，数詞，人名の場合: $k = 3$
- 検索語 q の品詞が名詞で，かつ，後続助詞が係助詞，格助詞，並列助詞の場合: $k = 3$
- 後続助詞として格助詞を取る名詞の掛かり先の動詞 q: $k = 3$
- 後続助詞として並列助詞を取る名詞の掛かり先の名詞 q: $k = 3$
- それ以外の名詞，動詞: $k = 1$
- それ以外の語: $k = 0.5$

(c) 両者の積 w_c

$$w_c(c,q) = w_o(c,q) \times w_p(c,q) \tag{5}$$

3. 評価実験および結果

3.1 辞書逆引き法の単体性能

以下のクロスワードを用いて，辞書逆引き法の単体性能を評価する実験を行った．

(1) 2012 年 9 月–2013 年 2 月，2013 年 6 月–11 月の『朝日新聞』(名古屋版) に掲載された 41 問，カギ 1,126 個
(2) 2004 年，2006 年，2008 年，2010 年，2012 年の『JAF Mate』に掲載された 47 問，カギ 1,465 個

この実験では，これらの問題に含まれるカギのうち，穴埋のカギを除いた計 1,913 個のカギを使用した．性能評価には，以下の 3 つの指標を用いた．

Recall(*N*): 候補リストの上位 *N* 件中に正解を含む割合
MRR: Mean Reciprocal Ranking: 正解順位の逆数の平均値
平均スコア: 正解ワードのスコアの平均値

実験結果を表 2 に示す．比較対象のベースラインには，出現頻度に基づく重み付け (TF-IDF 法) を採用した．この表において，提案手法は上記の 3 つの指標のすべてにおいて，ベースラインよりも良い値を示している．このことから，提案手法は TF-IDF 法よりも優れていることがわかる．

3.2 システムの全体性能

提案手法のうち，どの重みを採用すれば良いかを調べるために，辞書逆引き法を組み込んだクロスワードソルバー全体の性能を評価する実験を行った．このソルバーは，カギから答を推測する 17 種類の方法 (辞書逆引き法はそのひとつ) とグリッドを埋めるプログラムから構成されている．実験では，前述したクロスワード計 88 問を使用し，ソルバーが出力した解のワード正解率を調べた．

実験結果を表 2 のワード正解率の列に示す．この表より，使用する重みを変更しても，ワード正解率にそれほど大きな差はない．しかしながら，3 種類の重みの中では，『朝日新聞』と『JAF Mate』のどちらに対しても，$w_c(c,q)$ のワード正解率が高い．この結果に基づき，我々は，出現文節による重みと後続助詞による重みの積 w_c を，辞書逆引きの重み決定法として採用する．

この表に示した『朝日新聞』に対するワード正解率 89%は，先に報告した正解率 68% [1] より，大幅に向上している．この性能向上には，本論文で述べた辞書の逆引き法の改良以外に，連想語検索の導入や実例データベースの増強が寄与している．現在のソルバーの能力は，我々が目標とした「人間並み」というレベルにかなり近づいている．

参考文献

[1] 内木 賢吾, 佐藤 理史, 駒谷 和範：日本語クロスワードを解く：性能向上の検討, 第 27 回人工知能学会全国大会, 2013

[2] 内木 賢吾, 佐藤 理史, 駒谷 和範：日本語クロスワードパズルのカギの解法, 情報処理学会全国大会 講演論文集, Vol.74, No.2, 3R-5, pp.267-268, 2012.

[3] 西村 徹郎, 橋本 泰一, 徳永 健伸：単語の定義文による辞書検索, 言語処理学会第 12 回年次大会予稿集, pp.329–391, 2006

[4] 粟飯原 俊介, 長尾 真, 田中 久美子：意味的逆引き辞書『真言』, 言語処理学会第 19 回年次大会発表論文集, pp.406–409, 2013