単語と漢字の連想に基づく四字熟語の創作支援

中山 恭輔† 藤井 敦‡
† 東京工業大学工学部情報工学科
‡ 東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

日本語において伝統的に用いられている表現形式として四字熟語がある。四字熟語は、短い表現であるにもかかわらず、構成漢字から連想される意味を込めることができる。そこで、何らかのメッセージを伝えるための標語や世相を言い表すキーワードとして用いられることが多い。

本研究は四字熟語を自動的に創作することを目的とする。四字熟語の創作を通して、人間が造語を行う仕組みや意味を連想する仕組みを解明することを展望している。人手で創作された四字熟語の創作パターンをモデル化する事で、四字熟語の自動創作を実現する。人手で創作された四字熟語として住友生命の創作四字熟語[1]を参考にする。

住友生命の創作四字熟語では世相を表す説明文と既存語の発音を利用して四字熟語の創作が行われている。例えば「山死水迷」という創作四字熟語は「自然の崩壊」という世相を表す語句と「山紫水明」という四字熟語である既存語の発音を利用して創作されている。

四字熟語の創作パターンは複数ある。例えば、四字熟語ではない「鉄腕アトム」という既存語を元にした「撤湾跡夢」、「獅子奮迅」という既存語と発音があまり類似していない「知事文人」、「波乱万丈」という既存語の意味を含む「波乱盤上」がある。以上の例から、創作パターンは「既存語が四字熟語であるか」、「既存語と発音が類似するか」、「既存語の意味を含むか」の 3 つの項目においてそれぞれ 2 値の組み合わせで、2^3 通りのパターンがある。

本研究では、四字熟語に含まれている漢字から表現する内容を連想できるようにする。そのため、表現する内容と関係がない既存語の意味は含めない。また、創作する四字熟語の発音を既存語の発音に近づけることで、四字熟語を想起しやすくなる。そのため、発音から既存語を連想できる四字熟語を創作の対象とする。創作四字熟語では既存語として、四字熟語でない語も用いられている。しかし今回は、既存語を四字熟語に限定して四字熟語の自動創作を行う。

[1]http://cam.sumitomolife.co.jp/jukugo/

2 関連研究

本研究の目的は四字熟語の造語であり、短い語句の自動生成に関連する。具体的には、フレーズ生成 [3, 5]、隠語生成 [6]、造語 [1, 2, 4, 7] がある。

既存語を用いたフレーズを生成する研究には幾島ら [3]、松平ら [5] がある。フレーズを生成する研究はキーワード抽出や入力された単語の類義語を利用し、対象を表す新たな表現を生成する点において本研究と類似する。しかし、本研究が単語の生成を目的にするのに対し、フレーズを生成する点で本研究と異なる。

隠語生成を行う研究には木村ら [6] がある。隠語生成は隠語が表す対象の発音や対象に含まれている単語の類義語、表記上の類似性を用いている。本研究も既存語との発音の類似を考慮した創作を行う。しかし、対象が表す意味を考慮した創作を行う点で本研究は隠語生成と異なる。

本研究と同様に対象が表す意味や発音を考慮した研究には、柴田ら [7]、黄ら [1]、皆川ら [2]、三浦ら [4] がある。柴田ら [7] で造語される頭字語はネーミング対象のテキストのキーワードの頭文字を用いている。しかし、頭字語からは対象を連想する事はできない。黄ら [1] は造語される翻字候補の発音と翻字候補に含まれる漢字から対象が連想できる意味訳型翻字である。本研究と異なる点は中国語の漢字と単語の連想を行っている事である。皆川ら [2] では、柴田ら [7] と同様にキーワードを収集し、ネーミング案を造語する。しかし、発音に関する評価は行っていない。三浦ら [4] で造語されるネーミングは音象徴のみから強いという印象を持つ語を造語するため、本研究と異なり表現する内容は無い。

3 四字熟語創作の計算モデル

3.1 概要

黄ら [1] は、翻字対象の印象と発音を保持した翻字候補を造語するため、「発音モデル」と「印象モデル」を利用して翻字候補を評価する。本研究も表現する内容の印象と既存語の発音を保持した四字熟語の創作を目的とする。そのため黄ら [1] で提案されている手法を四字熟語の創作に適用する。

本研究で提案する四字熟語創作の計算モデルを図 1 に示す。

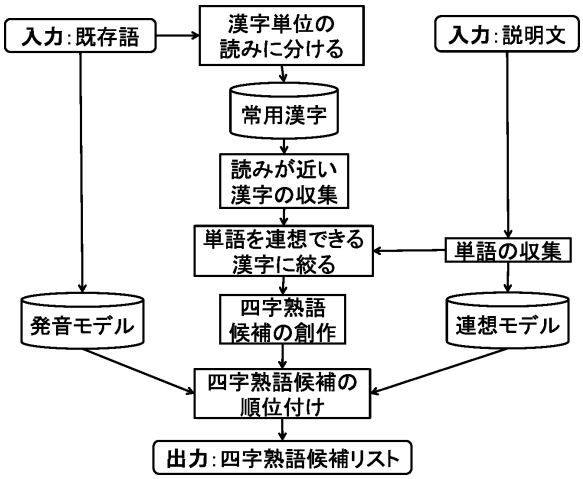

図 1: 四字熟語創作の計算モデル

図 1 に基づいて本研究で提案する四字熟語の創作支援手法について説明する。本手法の入力は説明文と既存語である。

説明文は創作される四字熟語が表す対象である。入力された説明文に含まれる単語を抽出し、その単語を連想する漢字を収集する。

一つ以上の既存語を入力し、創作する四字熟語の元になる発音として用いる。既存語に含まれる漢字の発音を利用して、四字熟語の創作に用いる漢字を収集する。1 章にあるように入力する既存語は四字熟語である。

入力された情報を手がかりに収集された漢字を用いて四字熟語候補の創作を行う。一般的でない漢字を創作に用いると四字熟語の意味が分からない場合がある。そのため、日常で用いる漢字として選ばれた常用漢字 2136 字を用いる。

四字熟語候補の創作を行う具体的な方法は、収集された漢字を 4 カ所に割り当てることである。しかし、自由に漢字を割り当てた場合、四字熟語中に同じ漢字が複数含まれる場合がある。本研究の目的から、異なる漢字を用いて入力する説明文の意味を最大限表現した四字熟語を創作することが望ましい。そのため、4 字の異なる漢字を用いて四字熟語の創作を行う。

創作された四字熟語候補を「漢字から表現する内容が連想できるか」と「発音から既存語が連想できるか」という観点から行う。さらに、本研究の創作パターンでは、説明文 d と既存語の発音 e は独立である。以上より、四字熟語候補 c を式 (1) によって評価する。

$$P(d, e|c) = P(d|c) \times P(e|c) \tag{1}$$

$P(d|c)$ と $P(e|c)$ はそれぞれ「連想モデル」と「発音モデル」に対応する。式 (1) によってスコア付けされた四字熟語候補を降順に出力する。

以下、3.2 節と 3.3 節でそれぞれ「連想モデル」と「発音モデル」について説明を行う。

3.2 連想モデル

表現する内容は、説明文で与えられる。説明文の意味は説明文に含まれる一つ以上の単語で構成される。説明文の意味を四字熟語で表すには四字熟語候補に含まれる漢字から、説明文に含まれる単語が全て連想されるのが望ましい。そのため式 (2) によって $P(d|c)$ を計算する。

$$\begin{aligned} P(d|c) &= \prod_{i=1}^{n} P(w_i|c) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{4} P(w_i|k_j) P(k_j|c) \end{aligned} \tag{2}$$

式 (2) において n は説明文 d に含まれる単語の数である。w_i は説明文 d に含まれる単語である。k_j は四字熟語候補 c に含まれる漢字である。$P(w_i|k_j)$ は漢字 k_j から単語 w_i を連想する確率である。$P(k_j|c)$ は四字熟語候補 c に漢字 k_j が含まれる確率であり、一様分布であると仮定すると $p(k_j|c) = \frac{1}{4}$ となる。

$P(w_i|k_j)$ は式 (3) を用いて計算する。

$$P(w_i|k_j) = \frac{F(w_i, k_j)}{\sum_{w} F(w, k_j)} \tag{3}$$

$F(w_i, k_j)$ は単語と漢字が共起する頻度である。共起頻度の計算には毎日新聞の 1991 年から 2011 年までの記事本文を用いた。

連想モデルは、人間が漢字からどのような単語を連想するかをモデル化する事を目的とする。人間が漢字から連想する単語は、漢字と意味的な関係で結ばれている単語である。例えば、「切」という漢字からは「ハサミ」という単語が連想できる。

本研究で用いたコーパスから単純に漢字と単語の共起頻度を計算する場合、漢字を含んでいる単語や文で頻繁に用いられている機能語や新聞記事で多用される単語の共起頻度が高くなる。その結果、漢字から連想できる単語が評価されないという問題が起こる。そのため、本研究では漢字と共起する単語をシソーラスを用いて収集し、収集した単語に対する共起頻度の計算を行った。

漢字と共起する単語を絞るために、日本語 WordNet[2]を用いて漢字から連想できる単語を収集する。日本語 WordNet は日本語のシソーラスである。日本語 WordNet は単語を類義関係のセット (synset) でグループ化している。さらに各 synset は上位下位関係など様々な意味的な関係で結ばれている。

漢字 k と意味上のつながりを持つ単語を探す手がかりとして漢字 k を含む単語を日本語 WordNet から探す。見つかった単語と同一 synset に属する単語、その synset の上位下位関係などの synset に属する単語、各 synset の定義文に含まれる自立語を漢字 k と意味上のつながりを持つと判断し、収集を行う。例えば「水」という漢字

[2]http://nlpwww.nict.go.jp/wn-ja/

表 1: 漢字が与えられたときの単語の連想確率

単語 w	$P(w\|$ 死$)$	単語 w	$P(w\|$ 山$)$	単語 w	$P(w\|$ 光$)$
葬儀	0.02762	前	0.01044	客	0.02005
事故	0.01790	人	0.00902	写真	0.01263
人	0.01743	行う	0.00719	人	0.01136
受ける	0.01303	見る	0.00716	見る	0.01093
前	0.01227	出る	0.00649	前	0.00985
思う	0.01192	世界	0.00599	思う	0.00787
出る	0.01166	自然	0.00585	目	0.00720
殺人	0.01119	問題	0.00578	使う	0.00693
移植	0.01039	入る	0.00555	行う	0.00671
行う	0.00945	言う	0.00538	出る	0.00602

に対して「ウォーター」、「湯気」、「飲む」等の単語が収集される。

synset の説明文を茶筌で形態素解析し、以下の品詞の単語を自立語として収集する。

> 未知語、名詞-一般、名詞-副詞可能、名詞-ナイ形容詞語幹、名詞-サ変接続、名詞-形容動詞語幹、名詞-固有名詞-一般、名詞-固有名詞-組織、名詞-固有名詞-地域、名詞-固有名詞-人名-一般、形容詞-自立、動詞-自立

また、収集された単語の中でひらがなだけで構成された「する」や「いる」などの単語で文字数が 2 文字以下の単語は機能語と判断して除外する。

以上のようにして絞った単語を用いて計算した連想確率を表 1 で示す。「死」に対して「殺人」のように人間が漢字から連想できる単語が含まれている。しかし、「前」や「人」のような一般的な単語も混ざっている。

3.3 発音モデル

四字熟語候補を「発音から既存語が連想できるか」という観点から評価する。四字熟語候補の発音から既存語を連想できるのは発音が類似するときである。発音の類似性は音素や音節の単位で発音が類似する確率を定義した音素や音節の Confusion Matrix を用いて計算することができる。しかし今回は実験的に、発音の類似性を測るために編集距離を用いる。編集距離を用いた $P(e|c)$ の計算は式 (4) を用いて行う。

$$P(e|c) = \frac{1}{Edit(e,h)+1} \tag{4}$$

式 (4) にある h は四字熟語候補 c の発音である。$Edit(e,h)$ は既存語の発音 e と四字熟語候補の発音 h の編集距離である。

本研究は編集距離を発音の類似性を測る尺度として用いる。そのため、文字ではなく音節に対して編集距離を計算する。また、単なる音節の編集回数で計算を行うと発音上は類似している音節も同じように計算される。そのため式 (5) のように修正を行う。

$$C[i,j] = \begin{cases} C[i-1,j-1] & (e[i]=h[i]) \\ D[i,j] + \min\limits_{i,j} C_a[i,j] & (e[i]\neq h[i]) \end{cases} \tag{5}$$

$$C_a[i,j] = \{C[i-1,j], C[i,j-1], C[i-1,j-1]\}$$

式 (5) にある $D[i,j]$ は行われた編集によって値を変える。通常の編集距離の計算では $D[i,j]$ の値は常に 1 である。本研究では $D[i,j]$ を式 (6) のように定義する。

$$D[i,j] = \begin{cases} 0.5 & (1.\text{ の編集を行ったとき}) \\ 1 & (\text{それ以外のとき}) \end{cases} \tag{6}$$

1. 濁音・半濁音・清音の入れ替え、拗音の挿入・削除、子音変化、母音変化、撥音・促音・長音の挿入・削除・入れ替え

1. の編集は住友生命の創作四字熟語において頻繁に行われていた編集である。そのため、これらの編集は他の編集に比べ発音が類似していると考え、式 (6) のように値を決めた。

一つの音節に対して 1. の編集が複数回行われる事がある。例えば「さ」が「ざん」になる場合は「濁音化」と「促音挿入」の 2 回の編集が行われる。複数回の編集が行われた発音は、1 回の編集が行われた発音よりも元の発音が異なっていると考えられる。そのため 1. の編集を 1 回行ったときよりも複数回行ったときに $D[i,j]$ の値が大きくなり、1. にない編集を行ったときよりも値が小さくなるように式 (7) で計算する。

$$D[i,j] = 1 - \prod_{i=1}^{n}(1-D_i) \tag{7}$$

式 (7) において n は編集が行われた回数であり、D_i は行われた編集に対応する値である。例外として、「子音変化」または「濁音・半濁音・清音の入れ替え」と「母音変化」が同時に行われる場合は発音が完全に変わるため $D[i,j]$ の値は 1 である。

4 実行例

4.1 実行方法

全ての常用漢字を用いて四字熟語を創作した場合、出力される四字熟語候補が膨大になり、現実的な時間で計算が不可能になる。そのため、評価される四字熟語が創作されるように漢字の絞り込みを行う必要がある。

具体的には、四字熟語候補を評価したときにスコアが大きくなるように、創作に用いる漢字を各モデルでスコア付けしてスコアが大きい漢字に絞って創作を行う。漢字を「連想モデル」と「発音モデル」でスコア付けをしたときにどちらのスコアを優先するのかをユーザの入力で決め、漢字のスコアを決める。

漢字のスコア付けは説明文に含まれる単語と既存語に含まれる漢字の発音ごとに行われる。スコアが上位である漢字を定数に絞って用いる場合、説明文にある単語の数によって創作に用いられる漢字が変化する。そのため、四字熟語の創作に用いる漢字の数 N をあらかじめ決め、単語の数 N_w として、スコアが上位 $\frac{N}{N_w \times 4}$ 件の漢字を用いる。今回は $N = 40$ として実行した。

表 2: 提案手法で出力された四字熟語の例

四字熟語	発音	$P(d,e\|c)\times10^{12}$
瞬田幽稲	しゅんたゆーとー	1.901
瞬田幽霜	しゅんたゆーそー	1.802
瞬麦幽稲	しゅんばっゆーとー	1.282
瞬麦幽霜	しゅんばっゆーそー	1.243
鬱暇昼糧	うっかちゅーろー	0.871

表 3: 「春夏秋冬」と発音が完全に一致する四字熟語の例

四字熟語	発音	$P(d,e\|c)\times10^{17}$
瞬荷秋湯	しゅんかしゅーとー	5.938
瞬火秋倒	しゅんかしゅーとー	0.980
瞬火秋頭	しゅんかしゅーとー	0.731
瞬掛秋湯	しゅんかしゅーとー	0.641
瞬火秋湯	しゅんかしゅーとー	0.562

今回、実行するときの入力として次の創作四字熟語を用いた。

- 創作四字熟語：瞬夏愁稲
- 説明文：束の間の夏が過ぎ、愁うべき稲の冷害。
- 既存語：春夏秋冬

入力に用いた創作四字熟語と提案手法で出力された四字熟語候補を比較することで考察を行う。

4.2 実行結果に対する考察

表 2 は「瞬夏愁稲」の説明文と既存語を元に提案手法によって出力された上位 5 件の四字熟語候補である。提案手法によって出力された四字熟語候補には「瞬夏愁稲」は含まれていなかった。

人手で創作された四字熟語に含まれる漢字と提案手法で出力された四字熟語に含まれる漢字を比較することで、連想モデルに対する考察を行う。「瞬夏愁稲」に含まれる漢字で、提案手法で出力された四字熟語に用いられている漢字は「瞬」と「稲」である。「瞬」は説明文に含まれる単語「束の間」を連想する漢字として用いられており、$P(\text{束の間}\,|\,\text{瞬}) = 1.019\times10^{-4}$ だった。表 2 にある漢字「暇」の $P(\text{束の間}\,|\,\text{暇}) = 6.377\times10^{-5}$ よりも値が大きい。また「稲」は「夏」を連想する漢字として用いられており、$P(\text{夏}\,|\,\text{稲}) = 2.084\times10^{-2}$ だった。提案手法で出力された四字熟語に含まれる「夏」を連想する「幽」の $P(\text{夏}\,|\,\text{幽}) = 6.544\times10^{-3}$ という値よりも大きい。以上から「連想モデル」は人が漢字から連想する単語をモデル化することができていることがわかる。

また、提案手法で出力した四字熟語に含まれていない「夏」と「愁」は説明文にある単語を日本語 WordNet で収集ができていなかった。しかし、説明文の中に「夏」と「愁」を含む単語がある。そのため、単語に含まれている漢字も創作に用いることができるように連想モデルを改良することが課題である。

発音に関しては、表 2 にある上位の四字熟語候補の発音が入力した既存語の発音と類似していない語が多い。また、出力された四字熟語候補全てについて既存語「春夏秋冬」と同じ発音である四字熟語はなかった。そのため、既存語の発音と完全一致にしたときに出力される四字熟語候補について考察する。表 3 のように既存語の発音と完全一致するように四字熟語を創作したとき、最も $p(d,e|c)$ が大きい四字熟語候補は「瞬荷秋湯」で $P(d,e|c) = 5.938\times10^{-17}$ だった。この値は表 2 の値と大きく異なる。このことから、四字熟語の評価で「発音モデル」の評価が「連想モデル」の評価に比べ過小評価されていることがわかる。「発音モデル」による評価が「連想モデル」による評価と同等になるように「発音モデル」の改良を行うことが今後の課題である。

5 おわりに

本手法は創作四字熟語に用いる漢字を絞るという対話処理により創作支援を行っている。しかし、現在の創作支援ではユーザが創作に関わる割合が少ない。そのため、ユーザがより四字熟語の創作に介在することで、より良い四字熟語が創作できるようにする必要がある。今回は「連想モデル」、「発音モデル」に対する評価を行っていないため、手法全体の評価と合わせて今後行う必要がある。また、「連想モデル」は漢字を用いたネーミングや未知語の意味を表記から推測することへの応用が可能である。

参考文献

[1] 黄海湘, 藤井敦. 中国語への翻字における関連語抽出の応用. 自然言語処理, Vol. 17, No. 2, pp. 3–24, 2010.

[2] 皆川恵理子, 藤井敦. 種々の造語法に基づく名付け親支援システム. 言語処理学会第 14 回年次大会発表論文集, pp. 512–515, 2008.

[3] 幾島克洋, 藤田篤, 佐藤理史, 横川睦, 岩本宜式, 片岡亮. HTML 文書からのリスティング広告の自動生成. 言語処理学会第 14 回年次大会発表論文集, pp. 504–507, 2008.

[4] 三浦智, 村田真樹, 保田祥, 宮部真衣, 荒牧英治. 音象徴の機械学習による再現：最強のポケモンの生成. 言語処理学会第 18 回年次大会発表論文集, pp. 65–68, 2012.

[5] 松平智史, 萩原将文. 対話型遺伝的プログラミングと電子化辞書を用いたキャッチコピー作成支援システム. 電気学会論文誌 C(電子・情報・システム部門誌), Vol. 125, No. 4, pp. 616–622, 2005.

[6] 木村友秋, 藤井敦. 評判情報の検索における隠語的造語法の応用. 言語処理学会第 15 回年次大会発表論文集, pp. 284–287, 2009.

[7] 柴田容子, 藤井敦, 石川徹也. 頭字語ネーミングの計算モデル. 言語処理学会第 12 回年次大会発表論文集, pp. 755–758, 2006.