コスト最小法を用いたことば遊び
～数字語呂合わせの自動生成システム～

青木賢太郎
慶應義塾大学大学院 政策・メディア研究科

Abstract: 形態素解析に用いられるコスト最小法で計算されるパスコストは文の文としての妥当性を表していると言える。そこで、本研究では各数字に割り当てられた単純な語呂から、長い数字列の持つ文としての妥当性をパスコストによってはかり、語呂合わせを自動生成するシステムを作成した。また、形態素解析器の参照する辞書の中身をあらかじめ数字列に置き換える事によって、入力された数字列を直接、形態素解析可能にし高速化を図った。

1 はじめに

語呂合わせの歴史というのは非常に古く、遡ればことばの読みそのものが意味や力を持つという言霊の時代にまで行きつく。そこまでいかないにしても、江戸時代の人々は好んで語呂を合わせ、歌や落語などの中でその巧さを競っていた。現代では、語呂合わせは主に何かを覚えるために使われている。百人一首、元素周期表、中国王朝などもあるが、我々の日常で一番良く使われるのが、歴史の年号、電話番号、円周率などの数字語呂合わせである。覚えるためだけでなく、覚えさせるためにも数字語呂合わせは使われている。例えば、企業は自社の電話番号をできるだけ意味の通った語呂合わせで申請し、祝日や記念日がそれに関連した語呂になっていることも多い。欧米の電話番号などが、アルファベットを割り当てたりして、なんとか意味をつけて覚えやすくしようとしているのに対し、日本では数字の読みの集合から別の語を連想することによって記憶するという方法が取られている。

数字語呂合わせは日本だけではなく中国語圏などでも行われている。例えば広東語で４は「セイ」と発音し、「死」と同じ音となり一般には、不吉な数字として忌み嫌われる。実は、これと同じ現象が古代日本で起きた結果を、我々は普段口にしている。我々は数字を「イチ、ニ、サン、**よん**、ゴ、ロク、**なな**、ハチ、**キュウ**、ジュウ」と数えることがあるが、これは呉音をベースとした読みに、「シ、ク」は「死、苦」の表現につながることから、「シ、シチ、ク」を和語読みの「よん、なな」と漢音の「キュウ」に置き換えたためである。[1]（表1 参照）

数字の語呂は現在でも縁起を担ぐために使われるが、やはり数字列を覚えるために使用されるのが身近である。我々が語呂合わせをする際、通常図１のようなプロセスを経る。

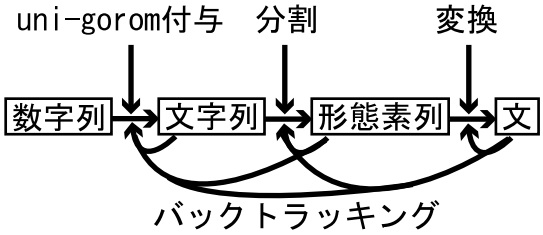

図 1: 語呂合わせのモデル

まず、数字列を各数字が持つuni-gramの語呂を用いて文字列に変換する。ここで、便宜上または語呂の良さからuni-gramの語呂のことを**uni-gorom**と呼ぶ。

続いてその文字列を形態素に分かち書きし、文として意味のあるものに変換するわ

けだが、その解析過程において何度もバックトラッキングが発生しているのが分かる。また、各過程において生成される組み合わせの総数は爆発的であり、その全ての経路を把握することは人間では不可能である。そこで本研究では、コスト最小法を用いた形態素解析によって、数字語呂合わせを実現するシステムを作成した。

2 関連研究

前節で述べたように、数字の語呂合わせは負担が大きく、またニーズも多いため、これを解決する方法はいくつか提案されている。しかしその何れも、自動生成というよりは、人間によるバックトラッキングの補助といった色合いの強いものである。

2.1 半自動システム

半自動のシステムには、WWW 上でCGIとして動作するものとスタンドアローンのアプリケーションとして動作するものがある。前者は主に携帯電話のインターネット接続サービスから、電話番号の語呂合わせをするものであり、後者は「ごろさく」[2]というソフトでこちらも電話番号を意識して12 桁までの入力となっている。「ごろさく」は語呂辞書に登録された3000 語以上の単語の組み合わせを効率良く提示できるが、CGI のものと同じく区切りはユーザが指定するようになっている。つまり多くの文節候補の中から、組み合わせを自分で選ぶという、最後の部分が手付かずである。

2.2 自動システム（プロトタイプ）

この他に、筆者による本システムのプロトタイプがある。本研究は言わばこれの改良手法にあたる。このシステムはコスト最小法に基づいて、数字列を解析する点では変わっていないが uni-gorom 付与の過程（図1）をそのままなぞっているため、最初に多くの文字列を生成していた。そのため、

- 中間ファイルが非常に大きい
- 時間がかかる（４桁で１分、５桁は待てない程）
- uni-gorom の語呂を分割してしまう（２７９→にしちく→西地区）

のような問題を抱えていた。

3 システムの特徴

本システムの要素技術は形態素解析に用いられるコスト最小法という技術であり、そのエンジンとして既存の日本語形態素解析器である ChaSen（茶筌）[3] を用いた。茶筌が参照する辞書の見出し語を予め数字列で置き換えておくことによって、直接数字列を解析することが可能になるわけである。そのため、数字列の桁数を気にせず高速な生成ができる。

3.1 語呂

辞書の生成に当たり、語呂の選定は非常に重要である。数字の語呂合わせにおいて、語呂が付与されるのは正確には数詞という品詞に分類される語である。その数詞は以下の様に分類される。[4]

- **基数詞** – 事物の数を表す語
- **序数詞** – 順序を表す語
- **倍数詞・分数詞** – 英語の double や half など

その他に集合数詞、配分数詞、不定数詞などがあるが、日本語では基数詞、序数詞が主であり、読みに対する語呂もこれらに付与されるのでここでは詳しく述べない。

3.1.1 uni-gorom

uni-gorom の生成ルールは以下の通りである。

- 数え方から（表1）
  - イチ、ニイ、サン – （漢語読み）
  - ヒイ、フウ、ミイ – （和語読み）
  - イー、アル、サン – （中国語）
  - ワン、ツー、スリー – （英語）
- 読み方から
  - 「読み」のうち全部、もしくは初めの 1 文字を使う
  - 濁音、撥音にも使ってよい
  - 長音でもよい
- 形から – エン＝円＝(0)、マル＝丸＝(0)、ワ＝輪＝(0)、オー＝英語読み＝(0)
- 意味から – ノー (0)、じ＝次＝(2)

また、数字の語呂合わせでは基本的には位取りは行わないが、100、1000、10000 については出現確率も高く読みも身近であるので、それぞれ「ひゃく」、「せん」、「まん」という uni-gorom を付与した。10 についても表1 の通りである。以上を組み合わせた結果、196 通りの uni-gorom が設定された。

3.1.2 50 音との対応

読みの最初の文字だけで、48 音中 33 音を表現した。その対応表が表 2 である。

3.2 コスト最小法

茶筌が用いているコスト最小法では、単語の出現のしやすさや、品詞（単語）のつながりやすさをコーパスなどから確率値として求め、それをコストとして辞書や文法に記述することによって、その最適パス（図 2 では太線）を求める。そのため、辞書と文法さえ揃えれば言語に依存しないで解析が行える。従って本システムでは表記が数字だけの言語を、形態素解析するのと同じことである。

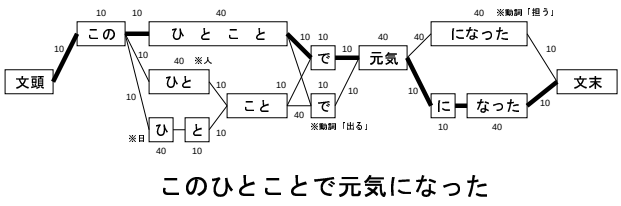

図 2: コスト最小法の例

3.3 辞書の作成

語呂合わせ用の辞書は以下の手順で作成した。

- 茶筌の辞書を uni-gorom のものだけにして、本来の辞書ファイルを解析
- 全ての uni-gorom の形態素コストを同一にし、品詞は名詞にすることで、uni-gorom が bi-gorom に優先して出現するようにした（例　れい　×れ＋い　○れい）

```
(品詞 (名詞 一般)) ((見出し語 (レイ 5)) (読み 0) )
(品詞 (名詞 一般)) ((見出し語 (イ 5)) (読み 1) )
```

図 3: uni-gorom 辞書

図3 において、表記と読みを入れ替えているのは、茶筌を用いて分かち書きと同時に変換を行うためである。表示のオプションで見出し語を表示せず、読みのみを表示することで変換が可能になる。[1] この uni-gorom 辞書を用いて本来のシステム辞書を語呂合わせ用の辞書に変換する。（図4）同様に活用定義ファイルも変更して初めて、数字の語呂合わせをすることが可能になる。

[1] ひらがなのみの対話コーパスを解析する際などに使われる Tips である

(品詞 (名詞 一般)) ((見出し語 (異例 2720)) (読み イレイ) (発音 イレイ))

↓↓↓↓↓

(品詞 (名詞 一般)) ((見出し語 (1 0 2720)) (読み 異例) (発音 イレイ))

図 4: システム辞書と変更後

以上の変更を行った結果、元の辞書の項目が23万あまりであったのに対し、語呂合わせ用の辞書では4万5千あまりと5分の1程になった。これは数字からは変換できない仮名が存在するためである。(表2参照) だが、「ごろさく」の3000語と比較すると15倍もの開きがあり、計算機が組み合わせを求める能力との差を歴然と物語る。

4 実行結果

実行結果を図5に示す。これは茶筌の解析結果をグラフィカルに表示するVisual Morphs[5]を使った例である。パスコストを見ながら自分の好きな経路を選べるようになっている。

5 発展

発展としては

- uni-gorom を選べる
- 概念ベースと組み合わせる

などが考えられる。前者では自分の馴染みの uni-gorom で解析することによって語呂を覚えやすくする効果が期待できる。

また、後者では予めキーワードを入力し、概念ベースを用いて類似度を計算するなどしてユーザの趣向にあった語呂を優先的に出力できるものと考えられる。(例 4129 – ブティックには「良い肉」よりも「良い服」など)

6 おわりに

本稿では、形態素解析に用いられるコスト最小法のアルゴリズムによって数字の語呂合わせを自動的に生成するシステムについて述べた。人間ではとても思いつかない量の語呂を生成し、それをパスコストという文の妥当性を示す指標で順位付ける事が可能になり、ユーザの負担が大幅に減少した。

今後はユーザの嗜好によるカスタマイズについて、また中国語版の辞書作成などについて検討したい。

謝辞

茶筌のTipsを教えてくださった石崎研究室博士課程の金子拓也さんに感謝致します。お陰様で道が開けました。またシステム全般において手伝ってくれた花川賢司君に感謝致します。

参考文献

[1] 三省堂編修所 編:ことばの知識百科, 三省堂,1995

[2] todome:ごろさく,http://hp.vector.co.jp/authors/VA001042/

[3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸:日本語形態素解析システム『茶筌』version 2.2.1 使用説明書,Nara Institute of Science and Technology,2000

[4] 小川芳男, 林大:日本語教育事典, 大修館書店,1982

[5] 松田 寛, 松本 裕治:品詞タグ付きコーパス作成支援GUI ツール Visual Morphs, 情報処理学会研究報告 2000-NL-137, p.98, June, 2000

音読み	0	1	2	3	4	5	6	7	8	9	10
呉音		いち	に	さん	し	ご	ろく	しち	はち	く	じゅう
漢音	れい	いつ	じ	さん	し	ご	りく	しつ		きゅう	
慣用音											じっ
訓読み											
		ひと	ふた	み	よ	いつ	む	なな	や	ここの	とお
		ひとつ	ふたつ	みつ	よつ	いつつ	むつ	ななつ	やつ	ここのつ	
				みっつ	よっつ		むっつ		やっつ		
	ゼロ	はじめ			よん		むい	なの	よう		

表 1: 数詞の日本語読み

ん	わ	ら	や	ま	は	な	た	さ	か	あ	*
ん	0	ら	8	0	8	7	た	3	か	2	a
*	ゐ	2	1	3	1	2	7	4	9	1	i
*	*	る	ゆ	6	2	ぬ	2	4	9	5	u
*	ゑ	0	8	め	へ	ね	10	7	け	8	e
*	0	6	4	も	ほ	0	10	3	5	0	o

表 2: 50 音対応表

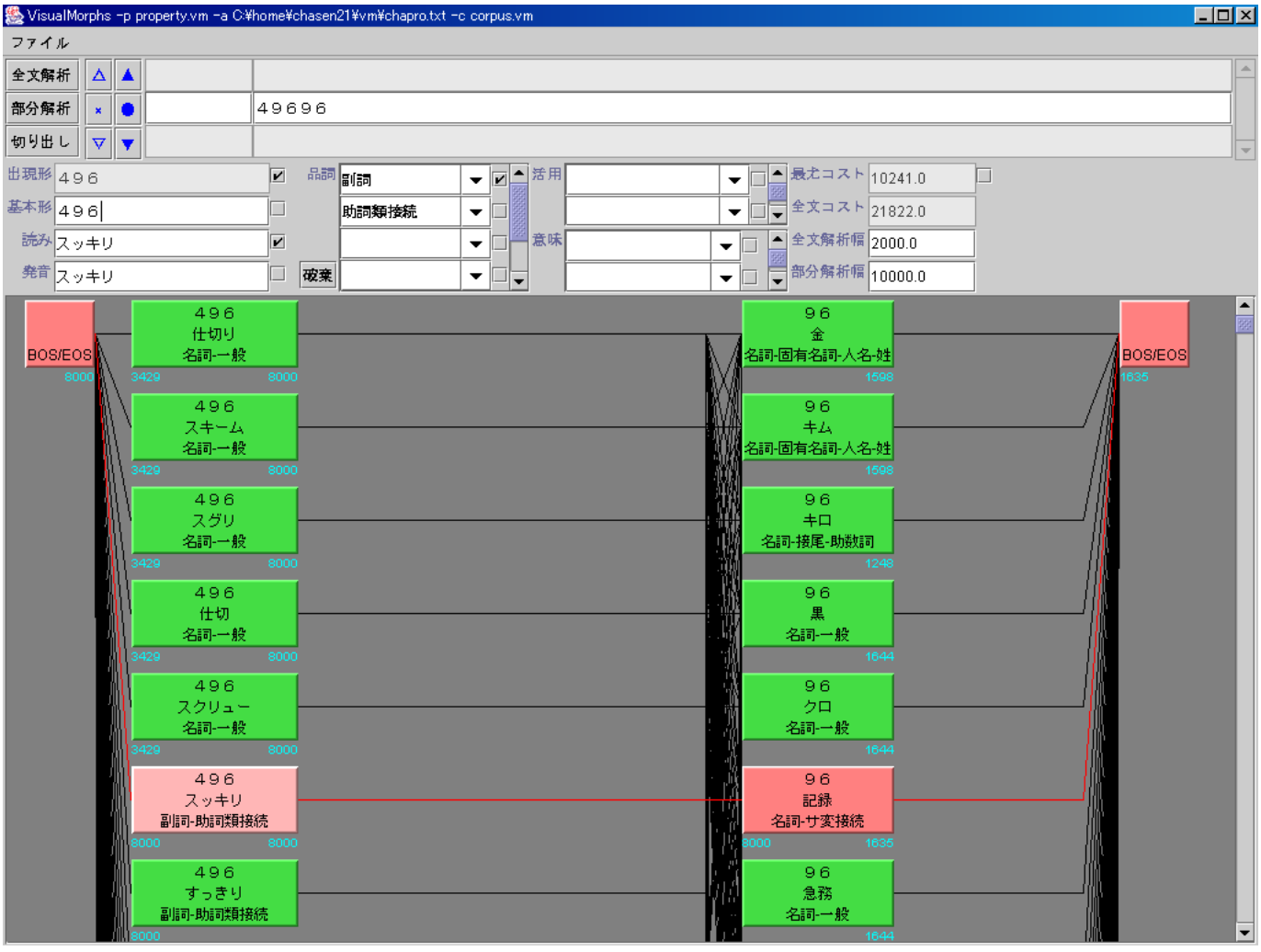

図 5: Visual Morphs を使った実行例