自然言語処理 117－5
（1997. 1. 20）

機械可読辞書の見出しについて

田中康仁
兵庫大学
yasuhito@humans－kc.hyogo－dai,ac.jp

概要
機械可読辞書の見出しをどのように選ぶかについて問題点を検討した。また機械可読辞書として既に存在している2社のものについて、日経コーパスからの機械的に抽出した2文字漢字列、3文字漢字列、カタカナ語をもちいてどの程度の頻度をカバーしているか調べた。また、2社の品詞付けの問題点、読みがな等についても調べ問題点をのべている。

Machine-Readable Dictionary Headings Professor

Yasuhito Tanaka
Hyogo University

Email: yasuhito@humans-kc.hyogo-dai.ac.jp

Abstract
We have examined the problems associated with the way in which machine-readable dictionary headings are selected. There are already products developed by two companies available as machine-readable dictionaries. Using the double-character and treble-character word sequences and the katakana words extracted by machine from the Nikkei Corpus we have investigated to what extent their frequency is covered. We have also studied the problems of accidence in terms of the away in which these two companies allocate parts of speech as well as problems such as yomigana (pronunciation-indicating kana).

〔0〕 はじめに

自然言語処理においてはいかに良い機械可読辞書を作るかが重要な課題である。

ここでは機械可読辞書の見出し語について考えてみることにする。いかに見出し語を集め選別し、その見出し語に情報を付け加えるかが重要である。これらについて考察した。

〔1〕 機械可読辞書の見出し語を集めるには

機械可読辞書の見出し語を集めるには次のようなことが基本となる。

(1) コーパスを分析し、その出現頻度の高いものを採用する。頻度の多さをどの程度で採用、不採用の基準にするかが重要である。

(2) 見出し語として短単位語を採用するか長単位語を採用するか考えなければならない。

一般に自然言語処理では語の曖昧さを減らすため長単位語を採用する。

長単位語を用いると単語数が増えるとか辞書の容量が増えるという欠点がある。しかし、曖昧さを減らせることを考えるとこれもやむをえない面がある。

(3) 語の中には一つのグループを作っているものがある。これらはグループに属する用語を全部入れたほうがよい。例えば、都道府県名、曜日、一月～十二月、太陽系等これらは一つの用語だけを入れて他を入れないと問題が起こる。

(4) 類語辞書、反対語辞書との照合

類語、反対語辞書と照合し、一方は入っているが、他方は入っていない等の矛盾が無いようにする。入れない場合は、その理由も明確にしなければならない。

(5) 機械処理用の辞書と人間が利用する本などの辞書とは考え方が異なる。

機械処理用の辞書と人間が利用する辞書とでは辞書に入れておくべき用語が異なることに注意しなければならない。

例えば人間では簡単に理解できるもの、また、用語の分割がわかるものでも、機械辞書としては入れておいたほうがよい場合が多い。

例　作業中、同月末

このような用語は人間の利用する本の辞書には見出し語としては採用されていない。

(6) 他の機械可読辞書との照合

機械可読辞書の原典が、人間が読んだり理解する辞書であれば、機械処理用の辞書としては不適合である。なぜなら人間は適当に理解し、推測できるが、機械処理では困難である。

(7) 見出し語の選択と削除

人が読んだり、理解したりするための辞書を開発した人の伝記等を読むと、集めた見出し語の全部を辞書の項目にするのではなく、その半分とか一部分しか利用していない。機械可読辞書についても同様の方法を考えなければならない。

また、時代と共に見出し語が古くなったり、使われなくなることもあるため、常に使われ方の頻度を分析しなければならない。使われ方と増減の測定が必要である。誤用、若者言葉、歌詞、ことわざ等も時代の流れの一つの産物である。ここに次のような資料を示す。

書　名	刊　年	新　規	削　除	収録語数
明解国語	1943	8,000	－	73,000
明解国語（改定）	1952	7,000	14,000	66,000
三省堂国語	1960	5,000	8,000	59,000
累　計		20,000	22,000	－

見坊豪紀「辞書をつくる」玉川選書 19pageより引用

(8) 表記のゆれについて

人間が利用する辞書では表記のゆれは一つに統一し、採用すればよいが、機械可読辞書では見出し語を増やして処理したほうが簡単に処理できることが多い。しかし、容量が増えるとか、この内容を印刷して使用すると項目が重複してよくないということもある。

〔2〕 データによる機械可読辞書との照合

データによる機械可読辞書の検討を行なった。二つの民間企業によって作成された機械可読辞書を用いた。

2-1) 三文字漢字列による照合

データは日本経済新聞データ5年分を用いた。このデータは日経5年分のコーパスから文字種の変わり目を利用して抽出した3文字漢字列である。

この3文字漢字列には誤って、また、たま

たま抽出されたものもあるので全数が一致することを望むものではない。

A社は大手メーカ数社と政府の出資によって作られた会社で、このプロジェクトは色々な機械可読辞書を作り出した。この中の日本語辞書を利用した。約２０万語が収録されている。

B社はワープロ・ソフトを商品化し、売り出した会社で、ワープロ・ソフトで一躍有名になった企業である。この企業の辞書７のものを利用した。約１３万語である。その後、この企業の辞書は改定され辞書１０までになっている。

次のような結果を得た。

異なりデータによるテスト

			%
総異なり数：		318,472	100%
A社内訳	完全一致数	11,157	3.50%
	時詞	8	
	副詞	9	
	一般名詞	10,272	
	固有名詞	211	
B社内訳	完全一致数	35,517	11.15%
	一般名詞	27,894	
	名詞形動	1,253	
	名詞サ変	1,572	
	固有地名	4,021	
	固有人名	369	
	固有一般	181	
	固有組織	218	
	数詞	3	
	副詞	6	

延べデータによるテスト結果

総頻度数		8,382,228	100%
A社内訳	完全一致数	2,891,570	34.50%
	時詞	21,398	
	一般名詞	2,775,161	
	副詞	1,404	
	固有名詞	93,607	
B社内訳	完全一致数	5,628,746	67.15%
	一般名詞	4,195,424	
	名詞形動	532,336	
	名詞サ変	330,599	
	固有地名	408,258	
	固有人名	34,143	
	固有一般	7,232	
	固有組織	119,994	
	数詞	534	
	副詞	226	

2-2) 結果のさらなる分析

不一致になったデータのうち頻度の高いものを３０ケずつ表示した。

A社の不一致データと頻度			B社の不一致データと頻度		
001	投資家	19,386	001	九二年	13,565
002	十五日	15,425	002	九一年	12,852
003	十八日	15,197	003	八九年	12,109
004	十九日	15,175	004	月連続	11,538
005	十七日	14,838	005	九三年	10,000
006	十六日	14,743	006	米政府	9,640
007	十二日	14,334	007	日午後	8,952
008	十四日	13,714	008	八八年	8,900
009	十三日	13,601	009	昨年末	8,579
010	九二年	13,565	010	面参照	8,217
011	十一日	13,316	011	日午前	7,408
012	九一年	12,852	012	八七年	6,661
013	八九年	12,109	013	八五年	6,284
014	月連続	11,538	014	位置付	6,194
015	高水準	11,488	015	九四年	5,875
016	日本側	11,411	016	九五年	5,736
017	来年度	10,724	017	年満期	5,371
018	九三年	10,000	018	三月末	5,252
019	売買高	9,941	019	八六年	5,151
020	新会社	9,653	020	三年間	5,086
021	米政府	9,640	021	五年間	5,013
022	報告書	9,312	022	年計画	4,593
023	大阪市	9,297	023	昨年秋	4,556
024	利用者	9,292	024	同協会	4,444
025	日午後	8,952	025	先週末	4,410
026	副社長	8,946	026	六月期	4,397
027	八八年	8,900	027	資金繰	4,374
028	昨年末	8,579	028	米国内	4,372
029	事実上	8,461	029	三年後	4,272
030	本格化	8,413	030	二年間	4,258

2-3) 三文字漢字についての二社の結果

A社の辞書は優秀な辞書ではあるが大量のコーパスによるデータの分析がなされていないようである。“投資家”、“高水準”といった用語が入っていないことも目につく。数詞についての取り扱いも十分ではない面が目につく。地名等の収集、採録も必要である。

B社は民間企業であるため、ワープロ等の商品を出しているため色々な検討がなされているので良い照合結果が得られたのであろう。今ではさらに良い辞書になっていると思われる。

この良、否の判断は、この論文の読者の判断にまかせたい。

2-4) 二文字漢字列による照合

三文字と同様に二文字漢字列を用いて実験を行いその結果は次のようになった。

異なりデータによるテスト結果

A社

品詞	異り
連体詞	2
時詞	53
副詞	120
一般名詞	33700
固有名詞	1009
合計	34884……37.49%
全データ	93047

B社

品詞	異り
固有一般	177
名詞形動	1256
数詞	67
名詞ザ変	1
固有地名	1579
副詞	74
固有人名	5107
一般名詞	17173
固有組織	301
名詞サ変	6252
合計	31987……34.38%
全データ	93047

延べデータによるテスト結果

A社

品詞	延べ
連体詞	63
時詞	152386
副詞	36997
一般名詞	23566787
固有名詞	1256767
完全一致数	25013000…… 93.27%
総頻度	26817980……100.00%

B社

品詞	延べ
固有一般	23138
名詞形動	1458717
数詞	94431
名詞ザ変	7
固有地名	948405
副詞	24033
固有人名	560425
一般名詞	11720223
固有組織	95392
名詞サ変	9722715
完全一致数	24647486…… 91.91%
総頻度	26817980……100.00%

2-5) 結果の分析

不一致になったデータのうち頻度の高いものを30ケずつ表示した。

	A社			B社	
001	上回	47,083	001	見通	80,838
002	相次	36,207	002	見込	75,040
003	下回	29,504	003	目指	68,801
004	万円	18,938	004	上回	47,083
005	億円	18,398	005	目立	43,268
006	一五	16,254	006	見直	37,275
007	二五	15,534	007	相次	36,027
008	役立	14,569	008	下回	29,504
009	都銀	12,843	009	値上	29,063
010	根強	11,698	010	万円	18,938
011	両社	9,946	011	一五	16,254
012	一三	9,669	012	手掛	15,581
013	同市	9,033	013	二五	15,534
014	月期	9,012	014	仕組	15,396
015	一四	8,933	015	値下	15,136
016	小動	8,712	016	役立	14,569
017	米側	8,477	017	見送	13,226
018	東芝	7,850	018	幅広	13,111
019	一七	7,812	019	利回	12,718
020	二年	7,803	020	一一	12,187
021	七五	7,682	021	根強	11,698
022	五年	7,341	022	三人	11,472
023	同省	7,302	023	一二	10,855
024	月物	7,170	024	利食	10,683
025	一九	7,054	025	間違	10,264
026	歯止	6,876	026	気持	9,924
027	二倍	6,723	027	軒並	9,678
028	長引	6,542	028	一三	9,669
029	同二	6,218	029	月期	9,012
030	約一	6,067	030	一四	8,933

2-6) 二文字漢字についての二社の結果

A社の辞書は固有名詞（人名、地名、組織名）等が極めて少ない。一般名詞が多いのが特徴である。サ変名詞の数が少ない。これは別の項目としてあるのかもしれない。B社より異り、延べ件数での比較で多い。数詞に対して考慮されてない。

B社の辞書は固有名詞も入り、全体としてバランスよく入っている。異り数はA社より少ないが頻度の高いものを重点的に入力しているのか、よくまとまっている。辞書の更新もさらに数回行っているので、今ではよい辞書になっているであろう。

2-7) カタカナ語についての照合結果

日本経済新聞社の5年分のカタカナ語にA社、B社の辞書と照合した。

日本経済新聞社のカタカナ語

	種類	延べ件数	
	194,576	7,826,426	100.00%
A社、B社の両方で一致	6,323	5,452,876	69.67%
A社だけで一致	8,904	373,679	4.77%
B社だけで一致	3,330	695,120	8.88%

A社のカタカナ語辞書で74.44%、B社のカタカナ語辞書で78.55%が一致している。B社は一致した種類では9,653であり、A社は15,227である。これはB社の用語の選び方が、使用頻度をうまく利用しているように考えられる。A社は全体的に使用が確立した用語を入れているため現在の用語に対しては注意がはらわれていないようである。これはカタカナ用語の選定にあたって、人のもちいる辞書を利用したためである。

2－8）見出し語と読みについて

見出し語には読みが付けられる。これらはほぼ一致するのであるが、照合してみると次のような相異があった。

1）日経の二文字にA社、B社で読みを振って両方の読みに相違のあったもの

	件数	延べ
読みが先頭から違う	273	33,438
どちらかの読みが1文字多い（読みの文字別個数をとって比較すると延べ相違文字数が1になる）	84	140,423
読みが1文字違う（じ、ぢなど）（延べ相違文字数が2になる）	10	1,396
読みが1文字違う（上記以外）（延べ相違文字数が2になる）	71	5,989
読みが2文字違う（延べ相違文字数が3か4になる）	46	7,139
読みが3文字以上違う（延べ相違文字数が5以上になる）	15	10,332
それぞれに複数の読みが振られており、増減および相違がある。	246	141,257
	745	339,974

2）日経の3文字語にA社、B社で読みを振って両方の読みに相違のあったもの

	件数	延べ
読みが先頭から違う	13	192
どちらかの読みが1文字多い（読みの文字別個数をとって比較すると延べ相違文字数が1になる）	5	3,978
読みが1文字違う（じ、ぢなど）（延べ相違文字数が2なる）	3	124
読みが1文字違う（上記以外）（延べ相違文字数が2になる）	18	1,755
読みが2文字違う（延べ相違文字数が3か4になる）	19	1,390
読みが3文字以上違う（延べ相違文字数が5以上になる）	5	203
それぞれに複数の読みが振られており、増減および相違がある	5	79
	68	7,721

これらの相異はA社,B社の振りがなの誤り品詞の付け方の相異によるものや、その結果送りがなのつけ方に寄るもの等がある。又、古文に出てくる用語の読みと現代的読み方がある。

一致しなかったものを調べてみるとA社のものは古い読み方が付いており、B社のものは現代的な読みである。

例　大商人　A社（おおあきんど）
　　B社（だいしょうにん）

大量のデータの中にはこの程度の相異は含まれているのは当然である。しかし、少しずつ修正してゆくことを期待したい。

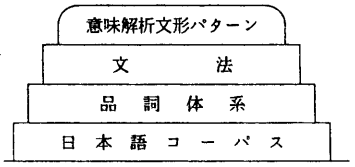
2－9）見出し語と品詞について

A社とB社では品詞体系に少し相異があるがあまり大きな相異はないものと思い次のようなマトリックスを考えてみた。

B社（2文字）品詞 ＼ A社（2文字）品詞	連体詞	感詞	副詞	一般名詞	固有名詞	合計 全データ
固有一般						
名詞形動						
数詞						
名詞ザ変						
固有地名						
副詞						
固有人名						
一般名詞						
固有組織						
名詞サ変						
合計 全データ						

しかし、品詞体系が少し異っていても大きな問題はないと思っていたが、思わぬところに異った品詞が付いていることが判った。これは大変なことである。B社は仮名漢字用の機械可読辞書であるが、自然言語処理用にも充分使えるものである。

品詞体系の上に文法があるのであるから品詞のゆらぎがあれば、文法にもゆらぎがあるのは当然である。文法が正しいものであるかを考えるにあたっては品詞付けが正しくなければならないのである。



［3］短単位語と長単位語の見出し語について

見出し語として長単位語を採用するか、短単位語を採用するかは大きな問題である。

短単位用語は主として次のような場合にもちいられている。

(1) 情報検索等の応用分野で使用される。

(2) 人間が利用する辞書

ｉ）照合率を向上させるためには短単位語をもちいる。

ii）辞書の冊数、厚さを増やさず、内容の充実と見出し語の増加のためになされる。

ｉ)、ii）では結果の分析や、語の意味の理解は人間が行うので短単位語が見出し語として利用される。

長単位語は主として次のような場合にもちいられている。

(1) 機械翻訳等の応用分野

(2) 意味処理を必要とするような自然言語処理

長単位語は専門用語などに使われている。曖昧性解消には有効である。しかし、辞書の容量が増大するとか、照合率（出現率）が低いといった問題がある。このため根気よく収集しなければならない。

しかし、長単位語から短単位語への分割は次のような方法で構造付分割データを持つことにより簡単に行える。

例　自然言語処理→（(自然・言語）処理）

←構造付の分かち書きを行っておくとよい。

しかし、{自然 言語 処理}　から自然言語処理には係り受けが一つに決まらない。幾つかの組合せが可能である。この中から一つのものに決定する一般的ルールを見つけるのは困難である。

［4］今後の課題

(1) 見出し語をいかに集めるかと同時に辞書項目（内容）としてどのようなものが必要か等を考えなければならない。これは質に関する問題でもある。

(2) 機械可読辞書を利用するにあたってはどのようにして開発されたものであるか、どのような特徴があるか、どのような性質や能力があるかも知るようにしなければならない。これはなかなか表示されていない問題である。見出し語数の多い、少ないだけでは計れないものがある。

(3) 人が語を索引し、読み、理解するために作成された機械可読辞書と機械処理用辞書と照合して何が不足しているか、何を余分に持たなければならないか調べてみる必要がある。

［5］おわりに

見出し語の選択は単純なことのようであるがこれが辞書の重要な点であることがわかった。コーパスを基にした頻度分析はコンピュータの力を借りるが、人間の少しの努力で大きな効果を得ることができる。

何を優先させたらよいかが明確に判る。判断のために使う労力を軽減させてくれる。しかし、コーパスの分析は定期的に行っておかなければ、ある時期の分析結果はすぐ古くさくなってしまう。人手によって先人が行った経験を体系化し計算機処理の中でどのように考えてゆくべきか検討すべきものである。計算機の機械的処理でコーパスの大量分析だけに頼るのではなく辞書の見出し語の検討、内容の検討がなされねばならないことがわかった。

機械可読辞書は数社からCD－ROMで販売されている。これらを参照しながらコーパスの分析も行いながら大規模な機械可読辞書の構築をめざすべきであろう。今後は100万語程度の見出し語の開発が望まれるであろう。

専門用語辞書の作成においても同じような問題を考えなければならない。そのためには専門分野のコーパスの調査と作成が重要である。

商品は多くの人々の批判を受ける中で改良され、良いものに発展するものであると確信している。

［6］データについて

この分析で用いたものは日経総合販売㈱から購入した「日本経済新聞CD－ROM 1990, '91,'92,'93,'94年版」である。

［7］参考文献

1）K．M．エリザベス・マレー　加藤知己訳「ことばへの情熱」上・下　三省堂出版　1984．5

2）見坊豪紀　「辞書を作る」　玉川選書　1976．11

3）西山里見とＱＱＱの会　「辞書がこんなに面白くていいかしら」ＪＩＣＣ（ジック）出版局　1996．6

4）田中康仁　機械可読辞書の更新について　自然言語処理112-17　情報処理学会　1996．3

二文字漢字列

A社

		時詞	副詞	一般名詞	固有名詞	連体詞	形容動詞	その他	合計
	固有一般	0	0	21	69	0	0	77	167
		0	0	702	1,517	0	0	4,876	7,095
	名詞形動	0	0	68	0	0	217	723	1,008
		0	0	166,614	0	0	284,544	889,973	1,341,131
	数詞	0	0	17	0	0	0	49	66
		0	0	51,958	0	0	0	42,320	94,278
B	固有地名	0	0	151	223	0	1	1,119	1,494
		0	0	16,585	654,235	0	2	218,926	889,748
社	副詞	1	44	2	0	0	9	18	74
		10	7,641	7	0	0	12,816	3,559	24,033
	固有人名	0	0	288	8	0	4	3,144	3,444
		0	0	14,956	816	0	137	117,033	132,942
	一般名詞	37	11	15,056	46	1	29	1,919	17,099
		140,809	15,821	9,671,375	112,185	61	2,943	1,648,404	11,591,598
	名詞サ変	0	1	5,924	0	0	0	229	6,154
		0	1	9,303,602	0	0	0	205,668	9,509,271
	固有組織	0	0	20	13	0	0	249	282
		0	0	7,191	27,379	0	0	53,085	87,655
	名詞ザ変	0	0	1	0	0	0	0	1
		0	0	7	0	0	0	0	7
	形容動詞	0	1	1	0	0	130	68	200
		0	2	380	0	0	46,215	11,612	58,209
	その他	12	63	9,832	376	1	226	52,548	63,058
		6,790	13,532	1,289,382	156,496	2	7,736	1,608,075	3,082,013
	合計	50	120	31,381	735	2	616	60,143	93,047
		147,609	36,997	20,522,759	952,628	63	354,393	4,803,531	26,817,980