Internal Use Only（非公開）

TR－SLT－0011

意味コード自動付与の品質評価
Evaluation of Quality about Automatic Addition of Semantic Codes

鷹尾和享
Kazutaka TAKAO

2002年3月 29日

概要

本稿では、外国語の知識を用いずに外国語の単語に対して意味コードの自動付与を行うための準備として、英語の単語で自動付与を試行し、付与性能の品質の評価を行った。対訳辞書の日本語訳を使って角川類語新辞典・日本語語彙大系を自動参照し、意味コードの自動付与を行った。実験セットには最近新しく英語意味コード辞書に加わった単語を用い、手作業で付与された結果を正解とし、それとの比較を行うことで品質を評価した。また、比較のため、日本語の単語に対しても同様の実験を行った。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288 京都府相楽郡精華町光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目次

第１章　概要 ……1

第２章　自動付与結果の品質評価方法 ……3

第３章　評価結果（英） ……4
3.1　集計結果 ……4
3.2　失敗例（角川） ……5
3.3　失敗例（語彙大系） ……7
3.4　失敗例（固有名詞） ……8

第４章　評価結果（日） ……10

第５章　今後の課題 ……12

参考文献 ……13

第１章

概要

筆者は韓国語または中国語の単語に対して意味コードの自動付与を行うことを目的とし、準備を進めてきたが、実際に韓国語・中国語に対して付与を行う前に英語および日本語で自動付与を試行し、付与性能の品質の評価を行った。筆者らは参考文献(1)・(2)で、対訳辞書の日本語訳を使って角川類語新辞典(3)・日本語語彙大系(4)を自動参照し、意味コードの自動付与を行う手法を提案したが、品質の評価を行うには至っていなかった。なお、参考文献とは違い、本稿では韓国語・中国語・さらには他の言語に対して自動付与を行うことを最終目標にしているので、外国語の知識に関しては基本的に利用せず、日本語の知識のみを使って意味コードの付与を行うこととした。すなわち、本稿で英単語に対して意味コード付与を試行する際、英日対訳辞書の日本語部分のみを使って辞書の参照等を行う。自動付与の方法の概略は以下の通りである。

(1) 既存の意味コード辞書を参照して自動付与（オプション）

上記の参考文献では表記のゆれ等を吸収して既存語の参照を行っていたが、本稿では異品詞の語の参照のみを行う。なお、既存の意味コード辞書がない場合や、利用したくない場合はダミーのファイルでよい。

(2) 角川類語辞典を参照

日本語訳で角川類語辞典を参照し、その意味コードを付与する。

(3) 語彙大系を参照し、マッピングを利用して角川に変換

日本語訳で日本語語彙大系を参照し、語彙大系→角川へのマッピングデータ（対応関係のデータ、参考文献(2)を参照）を利用して角川類語辞典の意味コードに変換する。対応関係がうまく取れているか否かによって"OK","要チェック"に分類できる。

(4) 何も参照せずに自力で付与

「○○症」「○○屋」等の特定のパターンの語の場合は、「○○」に入る語の種類が多く、辞書に載っていない場合があり、また、辞書を参照しなくても概ねわかるので、ヒューリスティックなパターンを幾つか用意して意味コードの付与を行う。なお、パターンの信頼性の度合いに応じて５段階

に分けることができる。

そのままの形で辞書にない場合は

・語形を変化させて参照

例：遅く　→　遅い

・途中で区切って切断点より後方の部分で参照（複合語）

例：アメリカンブレックファースト　→　ブレックファースト

・接尾辞を取り除いて参照

例：季節的　→　季節

等を行って辞書を参照した。このうち、「途中で区切って…」の場合は不適切な箇所で区切れて強引に辞書を参照する場合があるため、人手でチェックする必要のあるものを抽出する意味で別集計した。

なお、 http://lab.slt.atr.co.jp/~mtaka/jsem.html に記述があるように、角川類語辞典のアルファベットによる細分類は行わず、3桁の数字により分類を行う。

[開発環境]

Windows NT 4.0 SP5

Visual C++ 6.0

第２章

自動付与結果の品質評価方法

自動付与性能の品質の評価を行うため、以下の実験を行った。

[対象データ]

2000 年 7 月から 2001 年 11 月の間に英語意味コード辞書に新しく加わった語を意味コード付与実験の対象データとした。ただし、時期のずれ等の原因で英日対訳辞書に載っていない場合は実験対象から除外した。なお、対象データに含まれる語は主にフレーズブックの語であるが、それ以外の語も含まれている。

[正解データ]

上記の意味コード辞書に新しく加わった語は手作業で意味コードが付与されており、これを意味コードの正解とした。

[評価方法]

対象データの正解意味コードを伏せて自動付与を行い、その結果と正解データを比較した。品詞別・付与方法別に以下の尺度を用いて集計した。

exact = 正解に完全一致したエントリ数

extra = 必要なものは挙がっているが余計なものも挙がっているエントリ数

lack = 正しいものもあるが、不足もあるエントリ数

miss = 付与された意味コードはどれも不正解のエントリ数

含正解 = exact+extra+lack のエントリの割合、つまり少しでも合っていれば正解。

（最初から手作業のみで付与した場合と比べて手助けになったかどうかを表す）

再現 = 再現率の平均

再現率 ：正解数 /（正解数 + 不足意味コード数）

適合 = 適合率の平均

適合率 ：正解数 /（正解数 + 余分な意味コード数）

第 3 章

評価結果（英）

3.1 集計結果

表 1 ：集計結果（英）

		exact	extra	lack	miss	合計	含正解	再現	適合
合計		472	257	266	1752	2747	36%	30%	47%
品詞別	ADJ	37	13	40	74	164	54%	40%	50%
	ADV	4	3	9	20	36	44%	31%	43%
	CN	299	85	96	284	764	62%	55%	60%
	PROPN	113	143	82	1259	1597	21%	18%	40%
	V	19	10	36	34	99	65%	42%	51%
	PROPN 以外	359	114	184	493	1150	57%	47%	56%
付与方法別	角川	256	39	125	80	500	84%	68%	76%
	語彙:OK	29	26	15	40	110	63%	55%	41%
	語彙:要チェック	0	3	1	6	10	40%	35%	14%
	区切る:角川	61	37	42	205	345	40%	34%	29%
	区切る:語彙:OK	7	9	0	49	65	24%	24%	15%
	区切る:語彙:要ﾁｪｯｸ	0	0	0	6	6	0%	0%	0%
	自力(パターン 2)	1	0	0	1	2	50%	50%	50%
	自力(パターン 4)	5	0	1	1	7	85%	78%	78%
	付与できず	0	0	0	105	105	0%	0%	100%

集計結果を表 1 に示す。これを見ると、PROPN（固有名詞）の成績が際立って悪く、かつ、実験セットにおける PROPN の割合が高いため、全体の成績を悪化させていることがわかる。

3.2 失敗例（角川）

角川類語辞典で付与された語のうち、失敗例を幾つか挙げる。

※凡例

（"英語" 品詞 "角川コード"）　;<日本語訳の変形結果>　;;日本語訳

(1) 対訳辞書の日本語訳は１つだが、実際の語義は１つではない

[例 1]（"associate" V "411"）　;<連想>　;;連想する

自動付与＝ 411《思考》

正解＝ 220《離合》 267《総括》 290《関係》

[例 2]（"moldy" ADJ "257"）　;<黴びる>　;;黴びた

自動付与＝ 257《腐敗》

正解＝ 059《細胞》 139《新古》 145《匂い》 696《満足》 698《好き嫌い》

※英語の語義と、対訳辞書に書かれている日本語の語義は完全に一致するとは限らず、その意味で正解データを作った手作業付与担当者の努力の跡が伺える。

(2) 助動詞を取って辞書を引いたため意味が変わった

[例 1]（"close set" ADJ "215" "488"）;<近寄る>　;;近寄っている

自動付与＝ 215《接近》 488《親近》

正解＝ 223《接触》 226《並列》

[例 2]（"stabbing" ADJ "388"）　;<刺す>　;;刺すような

自動付与＝ 388《突き》

正解＝ 163《過激》 190《程度》

(3) 対訳辞書の訳語が不適切

[例 1] ("hipbone" CN "601")　;;;腰

自動付与＝ 601《胴体》

正解＝ 066《筋骨》

[例 2] ("landscaper" CN "040") ;;;景色

自動付与＝ 040《景色》

正解＝ 562《職人》

(4) 日英のニュアンスの違い

[例 1] ("patch" CN "392") ;;;継ぎ

自動付与＝ 392《修繕》

正解＝ 115《模様》 127《全部》 932《織物》

3.3 失敗例（語彙大系）

語彙大系と角川へのマッピングで付与された語のうち、失敗例を幾つか挙げる。

(1) 角川類語辞典に紛らわしい意味コードが複数存在する

[例 1]（"food poisoning" CN "608"） ;;;食中毒

自動付与＝ 608《病気》

正解＝ 073《発病》 074《生理》

[例 2]（"hochikisu" CN "961" "980" "989"） ;;;ホッチキス

自動付与＝ 961《筆記具》 980《工具》 989《刃物》

正解＝ 960《学用品》

(2) 語の用法の着目点の違い

[例 1]（"colourfast" ADJ "256"）;<色落ち> ;;色落ちする

自動付与＝ 256《色付き》

正解＝ 130《実質》

[例 2]（"overtime" ADV "103"）;<時間外> ;;時間外に

自動付与＝ 103《内外》

正解＝ 364《従業》 363《労働》

(3) 日英のニュアンスの違い

[例 1]（"unadon" CN "922"） ;;;うな丼

自動付与＝ 922《飯》

正解＝ 923《料理》

[例 2]（"recompression" CN "608"） ;;;潜水病

自動付与＝ 608《病気》

正解＝ 098《力》

3.4　失敗例（固有名詞）

固有名詞の失敗例を幾つか挙げる。なお、自動付与結果については

語彙大系の固有名詞意味属性　→　語彙大系の普通名詞意味属性　→　角川

のように自動付与の経路も表示した。

（1）地名

［例 1］（"akihabara" PROPN "725" "705"）;;;秋葉原

自動付与：

27 大字（その他）　→　0464 行政区画　→　705 《都道府県》

65 駅名等　→　0413 乗降場　→　725 《駅・港》

正解：705 《都道府県》 706 《都会》

※「駅」の抜けが補充された一方、語彙大系には都会かどうかの情報は載っていない。

［例 2］（"ama no hashidate" PROPN "507" "530" "581" "705"）;<橋立>　;;天の橋立

自動付与：

27 大字（その他）　→　0464 行政区画　→　705 《都道府県》

67 姓　→　0005 人間　→　507 《人》, 530 《仲間》, 581 《賢者》

正解：040 《景色》

※語彙大系は書き言葉への利用を主眼に作成されているため、駅名の「天橋立」は載っているが「天の橋立」の形は載っておらず、「橋立」で引かれてしまっている。また、語彙大系は観光地に関する固有名詞のカバレッジも弱いようである。

(2) 人名

[例 1] ("akemi" PROPN "511" "521" "523" "575" "578") ;;;アケミ

自動付与：

70 名（女） → 0049 女 → 511 《男女》, 521 《夫婦》, 523 《子》, 575 《俳優》, 578 《事務員》

正解：822 《名称》

※角川類語辞典には固有名詞は載っていないため、人名に 822 を付与するというルールは ATR で決めたものである。一方で、語彙大系－角川類語辞典のマッピングデータ作成には両者に共通して存在する語を使っているため、そのようなルールは恣意的に作成しない限り反映されない。また、語彙大系にかな書きのエントリがたまたま存在した例でもある。

(3) 団体名

[例 1] ("all nippon airways" PROPN "713" "729" "724") ;;;全日空

自動付与：

88 企業名 → 0374 企業 → 713 《団体》, 729 《店舗》
88 企業名 → 0428 仕事場 → 713 《団体》, 724 《仕事場》

正解：999 《航空機》 316 《運行》 713 《団体》

※語彙大系には企業名という認識はあるが、旅行会話で必要とされるような交通機関かどうかの語彙知識は載っていない。

第 4 章

評価結果（日）

英語との比較のため、日本語についても同様に実験を行った。なお、日本語については対訳辞書を持ち出すまでもなく、意味コード辞書に新しく加わった語を抽出するだけで日本語が得られるので、対訳辞書は不要である。

集計結果を表 2 に示す。日本語については普通名詞と固有名詞の区別がなく、普通名詞の中に固有名詞も含まれる。英語と同様、固有名詞については成績が悪いと思われるため、普通名詞の成績が悪くなっている。

一方、角川で付与された語は含正解・再現とも 97%で、非常に良い成績になっている。

表２：集計結果（日）

		exact	extra	lack	miss	合計	含正解	再現	適合
合計		4706	1486	379	3374	9945	66%	64%	63%
品詞別	サ変形容名詞	6	2	0	3	11	72%	81%	82%
	サ変名詞	375	34	11	86	506	83%	82%	81%
	形容詞	85	10	3	41	139	70%	69%	71%
	形容名詞	238	35	13	69	355	80%	78%	77%
	普通名詞	3142	1259	298	2753	7452	63%	61%	59%
	副詞	97	9	3	78	187	58%	57%	76%
	本動詞	622	133	50	254	1059	76%	73%	70%
付与方法別	既存の辞書	10	1	0	13	24	45%	66%	43%
	角川	3028	236	53	80	3397	97%	97%	93%
	語彙:OK	279	337	99	584	1299	55%	51%	35%
	語彙:要チェック	6	49	18	149	222	32%	28%	9%
	区切る:角川	1042	700	165	1119	3026	63%	60%	46%
	区切る:語彙:OK	106	114	40	547	807	32%	29%	20%
	区切る:語彙:要ﾁｪｯｸ	0	41	4	97	142	31%	32%	9%
	自力(パターン2)	77	0	0	0	77	100%	100%	100%
	自力(パターン3)	2	0	0	0	2	100%	100%	100%
	自力(パターン4)	31	8	0	15	54	72%	72%	64%
	自力(パターン5)	0	0	0	2	2	0%	0%	0%
	付与できず	125	0	0	768	893	13%	13%	100%

第５章

今後の課題

実験結果から、固有名詞に関する自動付与の性能が良くないことがわかった。一方で、日本語語彙大系は観光地等、旅行会話で必要となる固有名詞はあまり期待できないため、何らかの手段を考える必要がある。

また、本稿では英単語に意味コードを付与する際、対訳辞書の日本語訳を利用したため、英語１語に対して辞書参照に用いる日本語は１語となっている。英語の語義と日本語の語義は完全に一致するとは限らないため、市販の対訳辞書等を用いて、英語１語に対して複数の日本語を利用すれば自動付与の品質を改善できる可能性がある。

参考文献

(1) 鷹尾 和享, 柏岡 秀紀, 白井 諭: “異なる辞書を利用した意味コードの自動付与”, 情報処理学会第 59 回全国大会, 1N-07, (1999-9)

(2) Hideki Kashioka & Satoshi Shirai: “Automatically Expansion of Thesaurus Entries with a Different Thesaurus”, LREC 2000 (Second International Conference on Language Resources and Evaluation, Athens, Greece), pp.363-366, (2000.5.31-6.2)

(3) 大野晋, 浜西正人: “角川類語新辞典 CD-ROM 版”, 角川書店, (1989)

(4) NTT コミュニケーション科学研究所監修: 日本語語彙大系 全 5 巻, 岩波書店, (1997)

(5) Hideki Kashioka, Kazutaka Takao, Hiroko Ohta & Yoshiko Shirokizawa: “Applying TDMT to Abstracts on Science and Technology”, MT Summit VII (Singapore), pp.213-219, (1999.9.13-17)