専門用語コーパスにおける語彙的な階層付けの可能性

榎沢康子 (慶應義塾大学三田メディアセンター) suzume@mita.lib.keio.ac.jp
辻　慶太 (東京大学大学院教育学研究科) i34188@m-unix.cc.u-tokyo.ac.jp
影浦　峡 (学術情報センター--研究開発部) kyo@rd.nacsis.ac.jp

1. はじめに

学術情報センターでは，専門用語研究及び情報検索研究に向けた２種類のコーパスを作成している。１つは専門分野のテキストから成るテキストコーパス，もう１つは専門用語を集めた用語コーパスである。これらのコーパスは，専門分野における言語現象を，テキストと用語の両面から考察する為の基盤提供を目的としている[1]。用語コーパスでは，各専門用語は，国立国語研究所のβ単位に似た語構成要素（以下「語基」と記す）に分割してある。今後は，分野における中心的な用語，周辺的な用語を定義し，それらの別をコーパス中の各専門用語に付与する予定である。

今回は，そのための予備調査として，中心的・周辺的な用語を以下のように操作的に定義し，その特徴を調べてみた。即ち，ある分野のいくつかの「用語集」に共通して含まれる用語は，その分野において中心的な用語とみなす。本研究が分析対象とする人工知能分野・情報処理分野の用語は，後述するように，それぞれ３つの異なる情報源から収集している。ここでは，これら３つの情報源すべてに含まれた用語は，１つにしか含まれなかった用語よりも「中心的」な用語であると考える。その上で，中心的な用語にどのような特徴が見られるかを，主に語基の観点から調査した。調査項目は，語基の出現頻度，TFIDF，出現率の差に関する検定統計量，専門用語共通の語基を含む割合，語種の５つである。例えば，中心的な用語は，その分野に特徴的な語基を多く含むか，どのような語種から構成されるか，などを調べるのである。これにより，中心的な用語の特定に，語基の特徴を用いることの有効性を示し，また用語の中心性・周辺性を概念的に定義して行く上での基礎的資料が提示できると考える。

2. データ

本研究では人工知能分野・情報処理分野の用語を主な分析対象とする。まず人工知能分野については，『人工知能ハンドブック』[5]の索引から 1725 語（以下 AIHA），『人工知能大辞典』[6]の項目目次から 200 語（AIDC），索引から 3869 語（AIDI）を抽出し，これらをマージして重複を除いた人工知能用語コーパス AIT を作成した。情報処理分野については，『コンピュータ大百科』[7]の索引から 2362 語（IPCS），『情報処理用語大辞典』[8]の索引から 15031 語（IPDI），『英和コンピュータ用語大辞典』[9]の見出しから 31973 語（IPCM）を抽出し，同様の情報処理用語コーパス IPT を作成した。これらに含まれた用語・語基の数量は表１の通りである。なお以下では，AIT，IPT は「用語コーパス」，それらの元となった AIDC，AIHA，AIDI，IPCS，IPDI，IPCM の６つは特に「用語集」と呼ぶことにする。

表１．人工知能・情報処理の用語・語基数

		延べ	異なり
AIDC	用語	200	200
	語基	487	287
AIHA	用語	1725	1725
	語基	4294	1302
AIDI	用語	3869	3869
	語基	10027	2261
AIT	用語	5250	5250
	語基	13693	2706
IPCS	用語	2362	2362
	語基	5695	1671
IPDI	用語	15031	15031
	語基	36791	5145
IPCM	用語	31973	31973
	語基	84083	7517
IPT	用語	40082	40082
	語基	105024	9194

さて本研究では，含まれた用語集が多い語ほど，その分野で中心的な用語と定義する。例えば，AIDC・AIHA・AIDI のすべてに含まれた「知識表現」という用語は，AIDI １つにしか含まれなかった「自律無人飛行機」よりも，人工知能分野において中心的な用語と考える。情報処理分野についても同様である。

AIT の用語を，含まれた用語集の数で分け，１つだけに含まれた用語から成るコーパス AIT1，同様に２つ・３つに含まれた用語から成るコーパス AIT2・AIT3 をそれぞれ作成した。IPT についても同様である。それらの用語数は次のようになった。

表２．含まれた用語集の数に基づいて作成したコーパス

	用語数		用語数
AIT1	4766	IPT1	31542
AIT2	424	IPT2	7796
AIT3	60	IPT3	744

本研究は，上記３種類のコーパス中の用語を語基の観点から調査し，中心的な用語の特徴把握を目指すものである。

さて本研究では，後述する「共通語基」の特定，及び出現率の偏りの調査のため，以下の６分野の学術用語集もデータとして用いる。即ち，物理学[10]（以下 PHY），天文学[11]（AST），化学[12]（CHM），植物学[13]（BOT），農学[14]（AGR），建築学[15]（ARC），の６分野である。ここでは主に自然科学系の分野を選んでいる。人文・社会系分野も取り込んだ同様の研究は今後の課題としたい。

表3．学術用語集6分野の用語・語基の数

		延べ	異なり
PHY	用語	10636	10636
	語基	25119	4772
AST	用語	5119	5119
	語基	11784	2476
CHM	用語	11271	11271
	語基	22015	6427
BOT	用語	9314	9314
	語基	19498	5369
AGR	用語	15067	15067
	語基	29139	9103
ARC	用語	8480	8480
	語基	16718	5580

さらに，上記6分野と人工知能・情報処理分野の計8分野の用語をマージし，重複を除いた総合コーパス（以下GTと略す）を作成した。このコーパスに含まれた用語・語基の数量は次の通りである。

表4．GTの用語・語基の数

	延べ	異なり
用語	93242	93242
語基	221757	28322

3. 分析

上記のデータを用いて，語基の観点から中心的な用語の特徴を調査していく。ここでは次の5つの尺度を調査した。すなわち，語基の出現頻度，TFIDF，出現率の差に関する検定統計量，専門用語共通の語基の割合，語種の5つである。

3.1. 出現頻度

AIT1 ～ 3，IPT1 ～ 3 の各用語の語基に対して，それぞれ AIT・IPT における出現頻度を調べ，用語ごとに語基の平均頻度（以下 AF と略す），最小頻度（以下 MF と略す）を調べてみた。例えば AIT1 の「演繹 的 学習」という用語に対しては，AIT における「演繹」「的」「学習」の出現頻度を調べ，それら3つの平均頻度，3つのうち最小の頻度を調べるのである。AIT における各語基の出現頻度がそれぞれ 14，380，65 であったとすると，AIT1 における「演繹 的 学習」の AF は (14+380+65)/3=153，MF は 14 となる。結果は表5のようになった。この表から，例えば，AIT3 の用語の平均 AF は 42.36，IPT1 の用語の平均 MF は 35.42 であることが分かる。なお，B は各用語の語基数を表している。

表5．AF・MF の平均

		B	AF	MF
AIT1	平均	2.664	42.35	7.97
	分散	1.357	1949	126
AIT2	平均	2.059	37.95	14.78
	分散	0.718	1278	297
AIT3	平均	2.017	42.36	20.87
	分散	0.423	802	347
IPT1	平均	2.692	190.60	35.42
	分散	1.300	39265	2770
IPT2	平均	2.400	189.79	52.76
	分散	0.792	36120	5193
IPT3	平均	1.906	188.91	78.59
	分散	0.554	42203	12091

中心的な用語ほど平均 MF が高くなるのに対して，平均 AF はあまり変わらないことが分かる。この点については後述する。

3.2. TFIDF

上記の傾向に関連して，8分野の用語コーパス（PHY，AST，CHM，BOT，AGR，ARC，AIT，IPT）を用いて，次の TFIDF を各語基に対して調べてみた。この値は，キーワードの重要度を測る尺度として情報検索分野でよく用いられている[17]。

$$TFIDF=f*\log\frac{C}{D}$$

ここで，f：AIT（あるいは IPT）における語基 w の出現頻度，D：PHY，AST，CHM，BOT，AGR，ARC，AIT，IPT の8つのうち w を含んだコーパスの数，C：全分野数（この場合は8）である。

先ほどの「演繹 的 学習」を例にすると，「演繹」「的」「学習」を含むコーパスの数がそれぞれ，3，8，2 であった場合，AIT における各語基の TFIDF は，それぞれ 14*log(8/3)=5.96，380*log(8/8)=0，65*log(8/2)=39.13 となる。ここで先ほどの AF・MF と同様に，各用語の平均 TFIDF（以下 AT と略す），最小 TFIDF（同 MT）を次のように定義する。即ち，AT：用語に含まれた語基の，AIT（あるいは IPT）における TFIDF の平均（上記の例では AIT における「演繹 的 学習」の AT は (5.96+0+39.13)/3=15.03），MT：用語に含まれた語基の，AIT（あるいは IPT）における最小の TFIDF（上記の例では AIT における「演繹 的 学習」の MT は 0），と定義する。

これらに関して，AF・MF と同様の集計を行ったところ，次のようになった。

表 6 . AT・MT の平均

		AT	MT
AIT1	平均	5.562	1.003
	分散	58.7	6.00
AIT2	平均	8.875	3.117
	分散	123.6	49.65
AIT3	平均	14.434	5.857
	分散	184.5	109.52
IPT1	平均	31.499	5.157
	分散	1654.7	148.33
IPT2	平均	39.865	8.923
	分散	2195.8	340.76
IPT3	平均	45.519	18.080
	分散	3176.1	1477.96

AT の平均は，先ほどの AF と異なり，中心的な用語ほど高くなっている。MT についても同様である。ある分野での出現頻度が高くても，他の多くの分野に共通して現れる語基は TFIDF の値が低くなる。その意味で，上の結果から，中心的な用語ほど分野で特徴的な語基を含むと考えることできる。

3.3. 出現率の差に関する検定統計量

TFIDF は情報検索分野の尺度だが，次に統計的尺度も用いて調べてみた。即ち，AIT と GT（あるいは IPT と GT）における出現率（出現頻度をコーパスの全語基の出現頻度で割った値）の差について，次の検定統計量 Z を用いた。

$$Z=\frac{\frac{f1}{N1}-\frac{f2}{N2}}{\sqrt{a*(1-a)*\left(\frac{1}{N1}+\frac{1}{N2}\right)}}$$

ただし，f1：AIT（あるいは IPT）における語基 w の出現頻度，f2：GT における w の出現頻度，N1：AIT（あるいは IPT）における全語基の出現頻度，N2：GT における全語基の出現頻度，であり，a=(f1+f2)/(N1+N2) である。この Z は，標準正規分布に従う。

これまでの，AF・MF，AT・MT と同様に AZ・MZ を定義し，調べてみたところ，次のような結果になった。

表 7 . AZ・MZ の平均

		AZ	MZ
AIT1	平均	5.711	1.729
	分散	19.97	10.26
AIT2	平均	6.919	3.737
	分散	24.03	18.61
AIT3	平均	9.303	6.076
	分散	29.33	26.95
IPT1	平均	3.573	0.675
	分散	9.91	9.66
IPT2	平均	4.198	1.499
	分散	10.81	10.21
IPT3	平均	4.348	2.321
	分散	12.75	12.46

AT・MT と同じように，中心的な用語ほど AZ・MZ の平均が高いことが分かる。

ところで，ある語基 w の Z が± 1.96 内の場合，AIT と GT（あるいは IPT と GT）における w の出現率には，有意水準 0.05 で差が認められないことを考えると，IPT における平均 MZ の値はかなり低いと言える。平均 AZ に関しても，IPT の方が全体に AIT より低い。即ち，IPT は AIT と比較すると，語基の出現率に関して GT に近い性質を持っていると言える。AIT の用語数は 5250，IPT の用語数は 40082 と大きな差があることを考えると，上記傾向は，1 つの「分野」を構成する用語の集合に，規模の点で一定の限界があることを示しているように思われる。用語の中心性は「分野」の定義と密接に関わっている。上記の傾向を手がかりに，今後分野に関する分析も進めて行きたい。

3.4. 専門用語共通の語基

これまでの出現頻度の偏りに関する結果を補強する意味で，8 分野の用語コーパス（PHY・AST・CHM・BOT・AGR・ARC・AIT・IPT）すべてに共通して現れる語基を取り上げる。便宜上，以下ではこのような語基を「共通語基」と呼ぶ。これら共通語基が，用語中の語基に占める割合（以下 CR と略す）を AIT1 ～ 3，IPT1 ～ 3 について調べてみた。例えば，「演繹 的 学習」では全語基数は 3，共通語基は「的」1 つで，CR は 1/3 である。

表 8 . CR の平均

		CR
AIT1	平均	24.30
	分散	740.81
AIT2	平均	18.64
	分散	752.77
AIT3	平均	15.44
	分散	675.32
IPT1	平均	19.25
	分散	626.58
IPT2	平均	18.02
	分散	652.22
IPT3	平均	15.67
	分散	691.94

中心的な用語ほど，共通語基の占める割合が低いことが分かる。ところで，共通語基は全部で 239 個あり，これらの AIT，IPT における平均出現頻度は，それぞれ 15.64，95.41 であった。AIT における共通語基以外の語基 2467 個の平均出現頻度は 4.04，IPT における同様の語基 8955 個の平均出現頻度は 9.18 であった。このように，共通語基の平均出現頻度は，それ以外のに比べてかなり高い。先ほど出現頻度の節で，中心的な用語ほど MF が高いのに対し，AF はあまり変わらない結果を見た。これは，中心的でない用語ほど，高頻度の共通語基を多く含むこと，またその出現頻度がかなり高いため，用語の中で最小頻度に該当することが少ないためと考えられる。

3.5. 語種

最後に語種について調べてみた。表９で，w，k，g はそれぞれ和語，漢語，外来語の語基のみから成る用語を表し，wk，kg，wg はそれぞれ和語と漢語，漢語と外来語，和語と外来語のみから成る用語，wkg は和語，漢語，外来語すべての語基から成る用語を表す。また R は割合，S は個数を表す。この表から，例えば，AIT2 の用語において，漢語の語基だけから成る用語は 246 個あり，AIT2 全体の 58.02 %を占めること等が分かる。

表９．語種の割合

		w	k	g	wk	kg	wg	wkg
AIT1	R	0.59	47.61	11.31	10.47	24.95	0.63	4.45
	S	28	2269	539	499	1189	30	212
AIT2	R	0.00	58.02	18.63	5.19	16.75	0.47	0.94
	S	0	246	79	22	71	2	4
AIT3	R	0 00	55.00	23.33	0.00	21.67	0.00	0.00
	S	0	33	14	0	13	0	0
IPT1	R	0.97	30.11	21.31	8.60	33.51	1.72	3.78
	S	307	9497	6722	2713	10569	541	1193
IPT2	R	0.53	33.17	25.28	5.13	32.39	1.40	2.10
	S	41	2586	1971	400	2525	109	164
IPT3	R	0.54	33.20	40.73	3.09	21.91	0.40	0.13
	S	4	247	303	23	163	3	1

g の欄から，中心的な用語ほど，外来語の語基だけから成る用語が多いことが分かる。また w，wk，wg，wkg の欄から，全体に，中心的な用語ほど，和語の語基を含む用語の割合が低くなることが分かる。和語の語基には「の」や「な」といった助詞が多く含まれる。これらの語基は，用語全体に冗長な印象を与えるが，反面，他の語基がどのような関係で結合しているかを明示するため，用語全体の表意性を高める。中心的でない用語は，あまりその分野で一般的でないが故に，簡潔さを犠牲にして表意性を高めている用語が多いと考えられる。

4. おわりに

本研究では，人工知能・情報処理の２分野に関して，操作的に「中心的」な用語を定義し，それらの用語がどのような語基から構成されているかを調査分析した。その結果，中心的な用語ほど，その分野に特徴的な語基を含み，逆に「共通語基」を含まないこと，和語の語基を含む用語が少なく，外来語が多いこと，などが明らかになった。今後は，これらを手がかりに，また先述の「分野」の定義と併せて，用語の中心性に関する定義を考察して行きたい。その上で，今回の「共通語基」のように，語基にいくつかのタイプを設定し，同様の調査を行いたい。例えば，単独で１つの専門用語となれる自立的な語基，一般辞書にも収載されている日常語的な語基，などである。そのような調査を通じて，専門用語と一般語の境界も含めた，専門用語の質的・量的構造の一面が明らかになると思われる。

参考文献

[1] Kageura, K., Koyama, T., Yoshioka, M., Takasu, A., Nozue, T. and Tsuji, K. (1997) "NACSIS Corpus Project for IR and Terminological Research," *Natural Language Processing Pacific Rim Symposium 1997*, p.493-496.

[2] 宮島達夫 (1981)『専門用語の諸問題』秀英出版.

[3] 野村雅昭，石井正彦 (1989) "学術用語の量的構造" 日本語学, vol.8, no.4, p.52-65.

[4] 石井正彦 (1997) "専門用語の語構成：学術用語の組み立てに一般語の造語成分が活躍する" 日本語学, vol.16, no.2, p.21-41.

[5] 人工知能学会（編）(1990)『人工知能ハンドブック』オーム社.

[6] Shapio, S.C. and Eckroth, D. 大須賀節雄（監訳）(1991)『人工知能大辞典』丸善.

[7] Ralston, A. and Reilly, E.D. Jr. ed. 棟上昭男（監訳）(1987)『コンピュータ大百科』朝倉書店.

[8] 相磯秀夫（編）(1993)『情報処理用語大辞典』オーム社.

[9] コンピュータ用語辞典編集委員会（編）(1996)『英和コンピュータ用語大辞典』日外アソシエーツ.

[10] 文部省（編）(1990)『学術用語集：物理学編』学術振興会.

[11] 文部省（編）(1994)『学術用語集：天文学編』学術振興会.

[12] 文部省（編）(1986)『学術用語集：化学編』学術振興会.

[13] 文部省（編）(1990)『学術用語集：植物学編』学術振興会.

[14] 文部省（編）(1986)『学術用語集：農学編』学術振興会.

[15] 文部省（編）(1990)『学術用語集：建築学編』学術振興会.

[16] 海野敏 (1988) "出現頻度情報に基づく単語重みづけの原理" *Library and Information Science*, no.26, p.67-88.

[17] Salton, G. (1989) *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley.