機械翻訳用専門語句辞書の自動獲得

Automatic Acquisition of Translation Dictionary for a Nuclear Power Domain

楠見　好章　(Yoshiaki　Kusumi)*　松本　裕治　(Yuji　Matsumoto)†

要約　海外で発生した事象情報を原子力発電所の管理ないしは運転に反映し，安全の向上に役立てている．海外情報は英文テキストで入手するので，機械翻訳を活用している．しかしながら，翻訳品質はまだ良くない．

機械翻訳は文解析，構造変換，文生成を順次実行するが，構造変換において翻訳辞書を参照する．例えば，専門語句が文章に出現すると，専門語句辞書を参照し，適切な訳語を与えるしくみになっている．従って，現状では整備されていない専門語句辞書を整備することで機械翻訳の品質を向上させようと考えた．また，未登録語を逐一リストアップするのではなく，未登録語を自動的に獲得することをねらった．

本稿は，文対応済みの日本語と英語の対訳コーパスにおいて，日本語単語と英単語がそれぞれ出現する回数から二言語の表現の間の類似度を計算し，その計算結果から機械翻訳用専門語句辞書が獲得できるかどうか考察した．

キーワード　自然言語処理，機械翻訳，未知語，対訳コーパス

Abstract　A number of reports of overseas incidents have been utilized by utilities in managing and operating nuclear power stations to improve safety. Machine translation, one of the important applications of natural language processing, has been getting practical in use. When English texts are translated by machine translation software on the market, the translation quality is not satisfactory for practical use. Then, it is important to register words and phrases that are not included in the machine translation dictionary to enhance the quality of the system. If unregistered translation patterns are automatically obtained, translation dictionary will be effectively improved. A method of extraction of translation patterns from bilingual corpus and a translation system based on the extracted knowledge are discussed and evaluated with some experiments.

Keywords　natural language processing, machine translation,unregistered word, bilingual corpus

1.　はじめに

機械翻訳は文解析，構造変換，文生成を順次実行するが，構造変換において翻訳辞書を参照する．例えば，専門語句が文章に出現すると，専門語句辞書を参照し，適切な訳語を与えるしくみになっている．従って，現状では整備されていない専門語句辞書を整備することで機械翻訳の品質を向上させようと考えた．また，未登録語を逐一リストアップするのではなく，未登録語を自動的に獲得することをねらった．本稿は，文対応済みの日本語と英語の対訳コーパスにおいて，日本語単語と英単語がそれぞれ出現する回数から二言語の単語間の類似度を計算し，その計算結果から機械翻訳用原子力専門語句辞書が得られるかどうか考察した．

* (株)原子力安全システム研究所　技術システム研究所
現（社)海外電力調査会　電力国際協力センター
† 奈良先端科学技術大学院大学　情報科学研究科教授

2. 二言語の単語間の類似度計算

二言語の単語間の類似度計算概略を以下に示す．図１のように，まず，文対応した日本語テキストと英語テキストからなる対訳コーパスから，二言語の単語間の類似度を計算する．

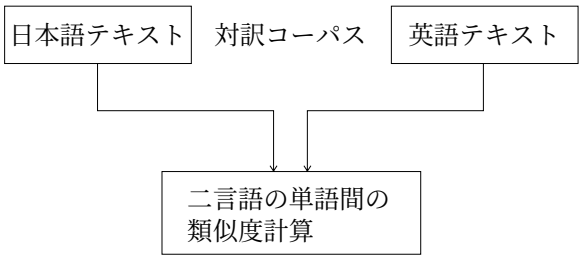

図1　二言語の単語間の類似度計算

二言語の単語間の類似度は，対訳コーパスが既に文単位で対応が付けられている場合，各単語（あるいは，単語列）がそれぞれ独立に出現する回数と対訳文に同時に出現する回数から求められる．ここでは，Dice係数を用いる従来法[1]を基に拡張した計算式を用いた．以下の式を対訳表現の類似度の計算式として定義した[2][3]．

$$\mathrm{sim}(<W_J, W_E>) = (\log_2 f_{je}) \frac{2f_{je}}{f_j + f_e} \qquad \text{(a)}$$

$$\mathrm{sim}(<W_J, W_E>) \geqq \log_2 f_{min} \qquad \text{(b)}$$

W_J ： 日本語単語（列）
W_E ： 英語単語（列）
f_j ： W_Jの出現回数
f_e ： W_Eの出現回数
f_{je} ： W_J，W_E対応の出現回数
f_{min} ： 出現回数の閾値

式(a) は，従来のDice係数に出現回数f_{je}の対数値で重み付けすることにより拡張した式である．これは，次の２つの理由による．

(1) 英単語2回,日本語単語2回,英日単語同時２回出現も,英単語100回,日本語単語100回,英日単語同時100回出現類似度は同じ１となるが,出現回数100回の方が2回に比べて信頼性が高いと考えられる．そこで出現回数を重み付けした．

(2) 計算量の削減である．このシステムでは，2回以上出現する１語以上の任意長の単語列を対象としているので，すべての計算を行うと計算量の爆発がおこる．それを避けるために，日英の単語が共起して現れる出現回数に閾値を設定した．上式は対訳表現の共起頻度に敏感な式になっている．

3. 実験及び考察

二言語の単語間の類似度計算結果について報告し，考察する．

3.1 実験内容

実験で使用した文献数は220例で，米国原子力発電運転協会（INPO）の原子力発電情報ネットワーク（NN：Nuclear　Network）を通じて入手しているNN情報及び世界原子力発電事業者協会（WANO）を通じて入手している事故速報（ENR），事故詳報（EAR）等の日英テキスト情報である．文数としては，約2,000文である．これらについて，英語，日本語それぞれを形態素解析し，単語間の類似度を計算した．

3.2 実験結果

類似度計算結果を表１から表６に示す．

表1　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
1	rod	棒	59	68	49	4.3326	20	◯
2	core	炉心	19	18	16	3.4595	10	◯
3	fuel	燃料	39	60	35	3.6268	10	◯
4	assemply	集合体	30	37	26	3.6481	10	◯
5	control	制御	50	68	43	3.9547	10	◯
6	RINGHALS	号機	18	21	16	3.2821	9	
7	thimble	案内管	13	13	12	3.3092	9	◯
8	test	試験	18	18	15	3.2557	9	◯
9	S shape	S	10	9	9	3.0031	8	
10	cycle	サイクル	11	12	10	2.8886	7	◯
11	step	ステップ	7	7	7	2.8074	7	◯
12	program	プログラム	6	6	6	2.5849	6	◯
13	refuel	取替	7	8	7	2.6202	6	
14	inch	mm	7	8	7	2.6202	6	
15	area	区域	8	7	7	2.6202	6	◯
16	be	れる	69	59	33	2.601	6	◯
17	drop	落下	11	13	9	2.3774	5	◯
18	OFF site power	外部電源	4	4	4	2	4	◯
19	inspection	検査	6	8	6	2.2157	4	◯
20	AFTER	後	6	9	6	2.0679	4	◯
21	17	17	4	4	4	2	4	◯
22	reactor	原子炉	13	14	9	2.1133	2	◯
23	turbine	タービン	7	7	6	2.2157	4	◯
24	10	K－10	6	5	5	2.1108	4	
25	fully	完全だ	7	7	6	2.2157	4	◯
26	time	時間	10	11	8	2.2857	4	◯
27	repair	修理	4	4	4	2	4	◯
28	vibration	振動	3	3	3	1.5849	3	◯
29	measure	測定する	4	5	4	1.7778	3	◯
30	maximum	最大	3	3	3	1.5849	3	◯
31	32	32	3	3	3	1.5849	3	◯
32	bank	バンク	3	3	3	1.5849	3	◯
33	monitor	モニタリング	6	4	4	1.6	3	
34	steam	蒸気	3	3	3	1.5849	3	◯
35	performance	性能	5	4	4	1.7778	3	◯
36	TRAINING	訓練	6	8	5	1.6585	3	◯
37	24	24	5	4	4	1.7778	3	◯
38	pressurized	米国	3	3	3	1.5849	3	
39	hole	穴	3	3	3	1.5849	3	◯

表2　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
40	position	挿入位置	9	5	5	1.6585	3	
41	4	4	12	21	9	1.729	3	◯
42	scram	トリップ	8	12	7	1.9651	3	◯
43	bowing	湾曲	10	22	9	1.7831	3	
44	speed	速度	3	3	3	1.5849	3	◯
45	drive	駆動装置	4	5	4	1.7778	3	◯
46	maintenance	保守	6	6	5	1.9349	3	◯
47	procedure	手順	5	4	4	1.7778	3	◯
48	august 22 1994	1994年8月22日	2	2	2	1	2	◯
49	34000 MWD	34,000MWd／	2	2	2	1	2	
50	be operate 51	出力51％運転中	2	2	2	1	2	
51	jam 18	18	2	2	2	1	2	
52	0.4429	内径	2	2	2	1	2	
53	house load	所内	2	2	2	1	2	
54	radiation dose rat	線量削減	2	2	2	1	2	
55	root cause	根本原因	2	2	2	1	2	◯
56	400　KV	400Kv送電	2	2	2	1	2	
57	low end	行程下半分	2	2	2	1	2	
58	following	取り上げる	2	2	2	1	2	◯
59	locate grid	６番目格子所	2	2	2	1	2	
60	recording are	録画する	2	2	2	1	2	
61	curve are	曲線	2	2	2	1	2	
62	work	作業	2	2	2	1	2	◯
63	nuclear	原子力工学	2	2	2	1	2	
64	engineer	技術者	2	2	2	1	2	◯
65	INFRARED	赤外線	2	2	2	1	2	◯
66	exposure	被曝量	2	2	2	1	2	◯
67	system	系	6	6	4	1.3333	2	◯
68	system	システム	2	2	2	1	2	◯
69	course	コース	2	2	2	1	2	◯
70	action	措置	4	5	3	1.0566	2	◯
71	parameter	パラメータ	2	2	2	1	2	◯
72	week	週間	2	2	2	1	2	◯
73	plant	発電所	8	8	5	1.4512	2	◯
74	periodic	定期	2	2	2	1	2	◯
75	equipment	機器	7	6	4	1.2307	2	◯
76	effective	する効果	2	2	2	1	2	
77	heat	熱	2	2	2	1	2	◯
78	tube	発生器	2	2	2	1	2	

表3　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
79	new	新	6	7	4	1.2307	2	◯
80	survey	調査	3	5	3	1.1887	2	◯
81	component	補	2	2	2	1	2	
82	trash	放射性	3	5	3	1.1887	2	
83	operation	運転	6	11	5	1.3658	2	◯
84	radioactive	放射性物質	2	2	2	1	2	
85	subsequent	その後	3	4	3	1.3585	2	
86	withdraw	引抜く	5	3	3	1.1887	2	◯
87	BINDING	固着	6	9	4	1.0667	2	◯
88	bow	湾曲	5	13	5	1.2899	2	◯
89	BURNUP	燃焼度	4	5	3	1.0566	2	◯
90	trip	トリップする	3	4	3	1.3585	2	◯
91	0	秒	9	5	4	1.1428	2	
92	DASHPOT	ダッシュポット	3	5	3	1.1887	2	◯
93	coolant	冷却材	3	5	3	1.1887	2	◯
94	48	48本	2	2	2	1	2	
95	16	16本	2	2	2	1	2	
96	natural	自然だ	2	2	2	1	2	◯
97	228	228	2	2	2	1	2	◯
98	investigation	調査	2	2	2	1	2	◯
99	reveal	判明する	4	3	3	1.3585	2	◯
100	project	プロジェクト	2	2	2	1	2	◯
101	annual	年１回	2	2	2	1	2	◯
102	OUTAGE	停止	5	13	5	1.2899	2	◯
103	friction	摩擦	3	5	3	1.1887	2	◯
104	old	古い	2	2	2	1	2	◯
105	additional	追加	2	2	2	1	2	◯
106	place	置く	3	4	3	1.3585	2	◯
107	affect	影響	3	5	3	1.1887	2	◯
108	15	15	2	2	2	1	2	◯
109	ultrasonic	超音波	2	2	2	1	2	◯
110	special	用特殊	2	2	2	1	2	
111	BY	よる	13	15	6	1.1078	2	
112	use	使用	13	10	5	1.0095	2	
113	checklist	チェックリスト	2	2	2	1	2	◯
114	change	変更	2	2	2	1	2	◯
115	folder	書	3	4	3	1.3585	2	
116	are	いる	16	38	9	1.0566	2	
117	unit	号	11	9	5	1.1609	2	

表4　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
118	jam	固着	7	5	4	1.3333	2	
119	also contribute	与える	2	3	2	0.8	1	
120	reduce PERSONN	放射線	2	3	2	0.8	1	
121	insert	上方停止する	4	3	2	0.5714	1	
122	insert	挿入する	2	4	2	0.6667	1	◯
123	K	動く　ない　なる	3	2	2	0.8	1	
124	jam	上方	3	3	2	0.6667	1	
125	next	次回	2	3	2	0.8	1	◯
126	prediction	予想	2	3	2	0.8	1	◯
127	verify	確認する	3	9	3	0.7924	1	◯
128	SECOND	増加	2	4	2	0.6667	1	
129	second	時間	2	3	2	0.8	1	◯
130	banana	バナナ	2	3	2	0.8	1	◯
131	include	含む	8	4	3	0.7924	1	◯
132	include	する　次　ある	5	2	2	0.5714	1	
133	contaminate	汚染	3	7	3	0.9509	1	◯
134	result	なる	7	13	3	0.4755	1	
135	job	おける	2	4	2	0.6667	1	
136	comprehensive	包括的だ	3	2	2	0.8	1	
137	core	及ぶ	3	2	2	0.8	1	
138	detect	発見する	5	4	2	0.4444	1	◯
139	hot	高温	3	2	2	0.8	1	◯
140	12	月	4	2	2	0.6667	1	
141	12	本	2	10	2	0.3333	1	
142	training	訓練	3	3	2	0.6667	1	◯
143	so	発生	2	6	2	0.5	1	
144	take	とる	4	2	2	0.6667	1	
145	ENGINEERING	技術部	3	3	2	0.6667	1	
146	process	プロセス	3	2	2	0.8	1	◯
147	trend	傾向	2	3	2	0.8	1	◯
148	day	日	3	4	2	0.5714	1	◯
149	other	他	5	2	2	0.5714	1	◯
150	other	位置	3	12	3	0.6339	1	
151	generator	チューブ	3	2	2	0.8	1	
152	4	タイプ	3	2	2	0.8	1	
153	year	年間	4	2	2	0.6667	1	
154	hige	高	4	2	2	0.6667	1	
155	hige	高い	2	2	2	1	1	
156	schedule	スケジュール　組む	4	2	2	0.6667	1	

表5　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
157	schedule	日程	2	2	2	1	1	◯
158	WITHIN	廃棄物	2	3	2	0.8	1	
159	diameter	制御棒　直径	4	2	2	0.6667	1	
160	diameter	11.25	2	2	2	1	1	
161	NONRADIOACTIV	非	2	3	2	0.8	1	
162	not	ない	5	9	3	0.6792	1	
163	guide	制御棒	12	13	4	0.64	1	
164	assembly	する　れる	4	19	2	0.1739	1	
165	grid	系	3	2	2	0.8	1	◯
166	BOWING	湾曲	3	8	2	0.3636	1	◯
167	transfer	電源	3	2	2	0.8	1	
168	there	冷却材	5	2	2	0.5714	1	
169	load	装荷する	4	2	2	0.6667	1	◯
170	insertion	挿入	6	11	4	0.9412	1	◯
171	cause	燃焼度	6	2	2	0.5	1	
172	cause	引起こす	4	2	2	0.6667	1	◯
173	event	事象	3	2	2	0.8	1	◯
174	0	径	5	2	2	0.5714	1	
175	increase	引上げる　られる	5	2	2	0.5714	1	
176	increase	落下	3	4	2	0.5714	1	
177	observe	ほど	4	2	2	0.6667	1	
178	3	現在	3	3	2	0.8	1	
179	design	配置	2	5	2	0.5714	1	
180	consist	各	2	6	2	0.5	1	
181	loss	喪失	2	3	2	0.8	1	◯
182	bottom	5	3	2	2	0.8	1	
183	BEFORE	前	4	6	3	0.9509	1	◯
184	start	起動する	4	3	2	0.5714	1	◯
185	OUTAGE	停止	2	6	2	0.5	1	◯
186	STARTUP	起動	2	2	2	1	1	◯
187	THE CONTROL	制御	2	19	2	0.1905	1	
188	last	装荷	3	3	2	0.6667	1	
189	low	小さい	5	4	2	0.4444	1	
190	BY	責任者	7	2	2	0.4444	1	
191	FOLLOWING	次	3	3	2	0.6667	1	◯
192	control	管理	7	10	4	0.9412	1	◯
193	PERSONNEL	要員	6	3	3	1.0566	1	
194	PERSONNEL	部	3	4	2	0.5714	1	
195	B.		3	24	3	0.3522	1	

表6　類似度計算結果

No.	W_E	W_J	f_j	f_e	f_{je}	sim(〈W_J, W_E〉)	出現回数の閾値(f_{min})	正解(◯)
196	be	する	34	130	16	0.7805	1	
197	be	ある	18	6	3	0.3962	1	
198	be	ない	15	6	2	0.1905	1	
199	A.		3	21	3	0.3962	1	
200	department	関する	3	4	2	0.5714	1	
201	use	利用する	8	3	3	0.8645	1	◯
202	use	使う	5	3	2	0.5	1	◯
203	individual	なる	3	10	2	0.3077	1	
204	are	可能だ	7	3	2	0.4	1	
205	operate	運転	3	6	2	0.4444	1	
206	2	2	7	10	3	0.5594	1	◯
207	2	2	4	2	2	0.6667	1	◯
208	unit	リングハルス	6	11	3	0.5594	1	
209	prior	前	3	3	2	0.6667	1	◯
210	C.		2	18	2	0.2	1	
211	rod	られる	10	13	3	0.4134	1	
212	management	管理	3	6	3	1.0566	1	
213	THIS	挿入	8	7	2	0.2667	1	
214	malfunction	実施する	2	11	2	0.3077	1	
215	2		3	16	2	0.2105	1	
216	also	られる	2	10	2	0.3333	1	
217	position	配置する	4	3	2	0.5714	1	◯
218	there	本	3	8	2	0.3636	1	
219	percent	リングハルス	3	8	2	0.3636	1	
220	finger	本	7	6	2	0.3077	1	
221	shutdown	停止	2	4	2	0.6667	1	◯
222	perform	実施する	2	9	2	0.3636	1	◯
223	FUELASSEMBLIE	燃料集合体	2	9	2	0.3636	1	
224	have	する	19	105	11	0.6138	1	
225	have	影響	8	2	2	0.4	1	
226	FUELASSEMBLY	燃料集合体	2	7	2	0.4444	1	◯
227	different	燃料	2	21	2	0.1739	1	
228	reactor	られる	4	8	2	0.3333	1	
229	1		3	14	2	0.2353	1	
230	3		5	12	2	0.2353	1	
231	are	られる	5	6	2	0.3636	1	

3.3 考察

類似度計算結果について，対訳関係にあり，そのままの形で翻訳用辞書として使用できるものを「正解」とし，人手により評価した．その結果，正解率は51.5％であった．ただし，例えば出現回数の閾値（f_{min}）が２以上の対訳表現の正解率は68.6％であった．出現回数の閾値（f_{min}）と正解率の関係を表７に示す．因みに，本実験で使用した文は，狭い分野に限定された統一的な内容を持った文であるため固定的な表現を含む類似の文が数多く存在した．そのため，構成単語数の異なる数多くの単語列候補が作成されてしまい，対応の候補が絞りにくい．例えば，「RINGHALS：号機」（表１　類似度計算結果）のような不適切な対訳表現が抽出され，正解率が下がったものと考える．また，通常の機械翻訳辞書には載っていないと思われる訳語，例えば「core：炉心」（表１　類似度計算結果）等が約2,000 文から10個程度得られた．

上記評価から，次の２点が言える．

(1) 図2　機械翻訳への適用図に示すように，文単位で対応が付けられている対訳コーパスを用いて，二言語の単語間の類似度を計算し，未登録語を自動的に獲得し，専門語句辞書を整備することで機械翻訳品質の格段の向上が期待できる．

(2) 特殊な専門用語を用いる分野の機械翻訳を行う場合についても，上記と同様の方法で専門語句辞書を整備することで，翻訳品質の格段の向上を期待できる．

表7　出現回数の閾値(f_{min}）と正解率

出現回数の閾値(f_{min})	正解率（%）
１以上	51.5
２以上	68.6
３以上	82.6
４以上	84.6
５以上	76.4

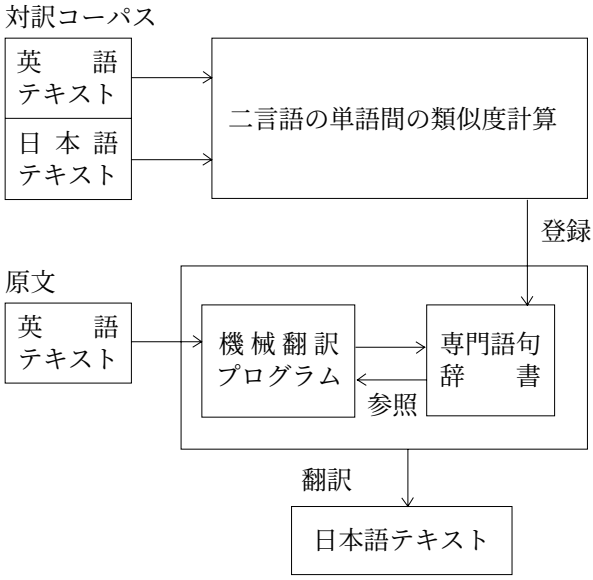

図2　機械翻訳への適用図

4. 結論

文対応済みの日本語と英語の対訳コーパスにおいて，日本語単語と英単語がそれぞれ出現する回数から二言語の単語間の類似度を計算した．その結果，現在，人手で作成している機械翻訳の専門語句辞書を自動的に獲得し，専門語句辞書を整備することで，機械翻訳品質の格段の向上が期待できる．

文献

(1) M.Kay and M.Röscheisen, Text-Translation Alignment, Computational Linguistics, 19(1), pp.121-142, 1993

(2) 北村美穂子，対訳コーパスからの翻訳知識の獲得と機械翻訳への適用，奈良先端科学技術大学院大学　修士論文，1995

(3) 北村美穂子　松本裕治，対訳コーパス中の共起頻度に基づく対訳表現の自動抽出，情報処理学会論文誌，vol.36, No.4, pp.727-736, 1997