修士論文発表

研究テーマ：CEFRレベルに基づく動詞-名詞コロケーションに関するコーパス研究

言語応用専攻 博士前期課程 2年
黄鈴娥
2017/2/8

発表の流れ

- 1. 研究の動機
- 2. 先行研究
- 3. 研究設問
- 4. 研究方法
- 5. 結果、考察
- 6. 結論
- 参考文献

1. 研究の動機

- 動詞 - 名詞コロケーション使用を調査し、名詞の分布を基準特性として、動詞や句動詞のレベルを機械的に分ける可能性を探求する。

2. 先行研究

- **2.1 語彙学習**

●語彙知識はコミュニケーションや他の言語スキルにおいて重要な役割を果たす。

- Celce-MurciaとRosensweig（1989）

学習の初期段階から受け入れられるべき

- Nation＆Waring(1997）

文法のように、語彙知識も重要

2. 先行研究

●語彙リスト

- The Dolch Word List (Dolch, 1948)
- The General Service List of English wordlist (see GSL, West 1953)
- the Academic Word List (Coxhead, 2000)

2. 先行研究

2.2 CEFRに基づく語彙レベル

2.2.1 CEFR

Proficient user	C2	Mastery
	C1	Effective Operational Proficiency
Independent user	B2	Vantage
	B1	Threshold
Basic user	A2	Waystage
	A1	Breakthrough

Adopted from Hawkins & Filipovic (2012, p.3)

2.2.2 English Vocabulary Profile

- ケンブリッジ大学
- CEFRに従う
- コロケーションや連語表現も含まれる

2.2.3 CEFR-J wordlist

- 日本の英語教育
- CEFRに基づく
- EVPより1レベル低い

2.3 コロケーションに基づく語彙レベル

2.3.1 コロケーションの定義

- 特定の意味や構文的制約
- Benson、Benson、Ilson（1986）2つのグループに分けられる：

文法コロケーション
語彙コロケーション

2.3.2 コロケーションにおけるSLA研究

- Nesselhauf（2005）
  頻度に基づく
- Benson(1990)
  語彙およびその教育的応用
- 近年
  コンピュータ技術　コーパスベース

2.3.3 コロケーションに用いる語彙難易度の判定

- Hill（2000）
  「語彙を知ること」の本質
  コロケーション使用など
- N. Ellis（2001）
  言語知識はコロケーション知識である。
- 単語の学習負担
  要因：単語を知ることの側面

What is involved in knowing a word

Nation (2013, p.40)

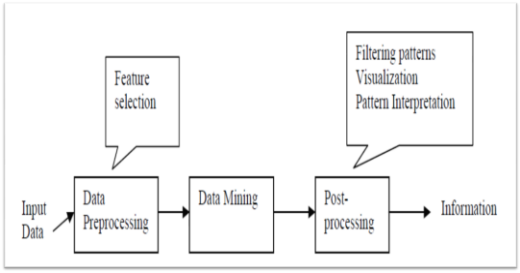
Form	spoken
	written
	word parts
Meaning	form and meaning
	concept and referents
	associations
Use	grammatical functions
	collocations
	constraints on use (register, frequency ...)

2.3.4基準特性としてのコロケーション

Hawkinsの仮説（2009）

	20 hypothesis by Hawkins
	1exical and grammatical areas for hypothesis testing
LEXICON	Lexical Choice Errors: Noun (N) and Verb (V)
	Lexical Occurrences: Noun (N) and Verb (V)
	Lexical Choice Errors: Adjective (J) and Adverb (Y)
	Lexical Occurrences: Adjective (J) and Adverb (Y)
	Verb Co-occurrence Errors
	Verb Co-occurrence Uses
	Infinitival Complement of Verbs: Errors
	Number Agreement NP-internally and on Verbs: Errors
SYNTAX	Missing Determiner Errors
	Determiner Choice Errors
	Word Order Errors: Verb-Object Separation
	Word Order Errors: Genitive Positioning
	Relative clause Uses and the Keenan-Comrie Noun Phrase Accessibility Hierarchy
	Relative Clause Errors and the AH
	Relative Clause Uses and the Link to the Subcategorizer (Verb/be+Adjective)
	Wh-movement Uses and the Link to the Subcategorizer (Verb/be+Adjective)
	Tough Movement Uses and the Link to the Subcategorizer (Verb)
	Raising Structure Uses and the Link to the Subcategorizer (Verb/be+Adjective)
OVERALL METRICS	Overall Error Counts
	Overall Syntactic Complexity Metric

2.4データマイニング



The Data mining process　　Adopted from (Tan, Steinbach, & Kumar, 2006)

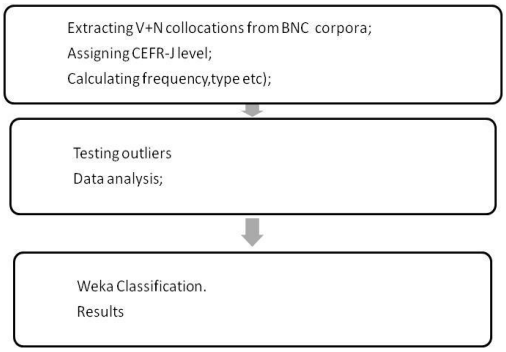
2.5 WEKA

- www.cs.waikato.ac.nz/ml/weka
- 機械学習ソフトウェア
- ランダムフォレスト

3. 研究設問

- RQ1動詞に共起する名詞のレベル、頻度と種類を基準特性として使用する場合は、コロケーション総頻度のみを使用するよりの分類精度が高いであるか？
- RQ2 11の名詞特性を使用するすべてのアルゴリズムの中で、どのアルゴリズムがTP/リコール率、FP率、精度、F-Measure、ROCエリアが最も優れているか？
- RQ3同じデータセット、特性、上記の実験から得られたパフォーマンスが最も高い分類器を使用して句動詞の予測結果は何か？

4. 研究方法



An example of the CEFR-J wordlist

headword	pos	CEFR	CoreInventory 1	CoreInventory 2	Threshold
'm	be-verb	A1			
're	be-verb	A1			
's	be-verb	A1			
a	determiner	A1			
a.m.	adverb	A1			
abandon	verb	B1			
abandoned	adjective	B2			
ability	noun	A2			
able	adjective	B1			
abnormal	adjective	B1			
abnormally	adverb	B2			
aboard	adverb	B1			
abolish	verb	B2			
aboriginal	adjective	B2			
aborigine	noun	B1	Nationalities and countries		

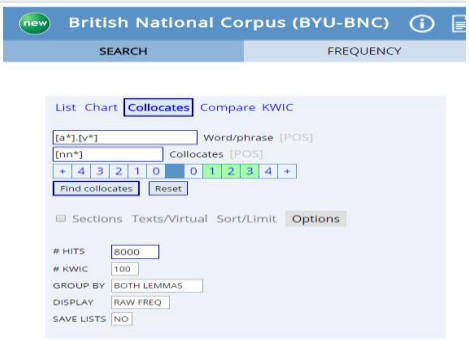
Total number of verbs in each CEFR-J level

CEFR-J levels of Verbs	Number of Verbs
A1	134
A2	205
B1	464
B2	546
Total	1349

Total number of nouns in each CEFR-J level

CEFR-J levels of Nouns	Number of Nouns
A1	632
A2	777
B1	1271
B2	1431
Total	4111

動詞ー名詞コロケーション抽出 BYU-BNCのWeb画面を使う



The samples of Verb-Noun collocations beginning with the letter "a"

SHOW TEXTS	CONTEXT	FREQ
1	[ABUSE] | | [CHILD]	77
2	[ABANDON] | | [IDEA]	64
3	[ABANDON] | | [PLAN]	61
4	[ABOLISH] | | [TAX]	51
5	[ABANDON] | | [ATTEMPT]	48
6	[ABUSE] | | [POSITION]	41
7	[ABUSE] | | [POWER]	40
8	[ABANDON] | | [POLICY]	32
9	[ABIDE] | | [RULE]	32
10	[ABANDON] | | [HOPE]	32
11	[ABSORB] | | [INFORMATION]	29
12	[ABANDON] | | [CLAIM]	29
13	[ABSORB] | | [ENERGY]	28
14	[ABSORB] | | [WATER]	28
15	[ABANDON] | | [YEAR]	26
16	[ABANDON] | | [CAR]	24
17	[ABANDON] | | [CHILD]	23

コロケーション総データ数

- 検索回数：470 (動詞数：1349)
- 毎回抽出データ平均数： 5000
- 総データ数：470 * 5000

CEFR-Jレベルを付与

No.	Verb	CEFR-V	Noun	CEFR-N	FREQ
1	abandon	B1	idea	A1	64
2	abandon	B1	plan	A1	61
3	abandon	B1	attempt	A2	48
4	abandon	B1	policy	B1	32
5	abandon	B1	hope	A1	32
6	abandon	B1	claim	B1	29
7	abandon	B1	year	A1	26
8	abandon	B1	car	A1	24
9	abandon	B1	child	A1	23
10	abandon	B1	project	B2	22
11	abandon	B1	principle	B1	21
12	abandon	B1	ship	A1	19
13	abandon	B1	position	A2	19
14	abandon	B1	pretence		18
15	abandon	B1	scheme	B2	17
16	abandon	B1	effort	A2	15
17	abandon	B1	practice	A1	15

基準特性

- levers of verbs (A1,A2,B1,B2)
- minimum frequency of nouns(A1minf,A2 minf,,B1 minf,,B2 minf)
- maximum frequency of nouns(A1maxf,A2 maxf,,B1 maxf,,B2 maxf)
- types of nouns(A1type,A2 type,,B1 type,,B2 type)
- total types
- total frequency

頻度と分布の統計

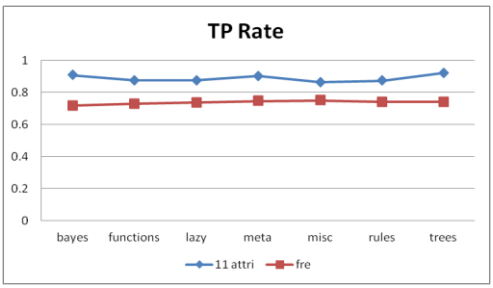
headword	CEFR	A1maxf	A1type	A1tfre	A2maxf	A2type	A2tfre	B1maxf	B1type	B1tfre	B2maxf	B2type	B2tfre
abandon	B1	64	139	758	48	95	419	32	103	456	22	52	180
abolish	B2	15	46	167	20	47	261	51	50	248	17	32	136
absorb	B1	29	57	287	20	78	350	23	65	259	28	27	120
accept	A2	81	248	2500	260	233	2662	290	314	3036	107	225	1568
access	B2	66	36	266	32	39	157	19	34	171	58	22	172
accompany	B1	40	140	728	33	105	415	17	105	367	19	62	199
accomplish	B1	7	22	78	23	10	50	9	11	39	9	6	21
account	B1	50	62	500	37	83	448	69	92	474	19	46	209
accumulate	B2	16	26	120	14	27	114	10	31	112	5	15	41
accuse	B1	15	72	256	159	44	443	28	58	265	28	47	230
accustom	B1	4	7	23	2	2	4	4	3	9	0	0	0
ache	B2	7	11	40	4	4	14	11	8	31	5	4	12
achieve	A2	327	181	2091	211	200	2438	424	234	2703	57	175	1076
acknowledge	B1	20	75	361	52	86	486	40	93	484	53	31	134
acquire	B1	162	155	1139	132	134	1301	81	133	867	30	101	447
act	B1	122	183	1421	131	162	1353	54	185	1044	49	162	945
adapt	B1	54	70	375	41	60	294	13	47	189	36	24	125
add	A1	141	331	3377	158	303	2821	180	341	2381	119	263	1402

5. 結果、考察

- RQ1

  動詞に共起する名詞のレベル、頻度と種類を基準特性として使用する場合は、コロケーション総頻度のみを使用するよりの分類精度が高いであるか？

  → YES



- RQ2

  11の名詞特性を使用するすべてのアルゴリズムの中で、どのアルゴリズムがTP/リコール率、FP率、精度、F-Measure、ROCエリアが最も優れているか？

  → Randon Forest

- **●TP/Recall率: RandomForest > BayesNet > RandomCommittee > KStar > MultilayerPerceptron > PART > JRip,**
- **●FP 率: BayesNet< RandomForest< RandomCommittee< PART< JRip< MultilayerPerceptron< KStar.**
- **●精度 RandomForest は一番高い.**

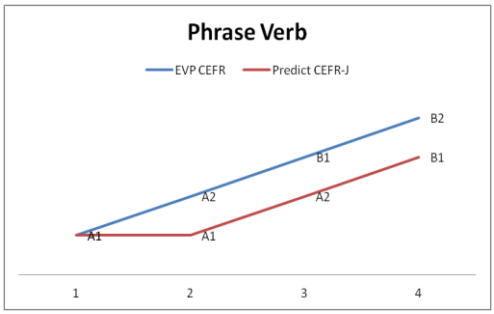
- **●F-Measure:RandomForest(0.923) > BayesNet > RandomCommittee > KStar >MultilayerPerceptron> PART > JRip.**
- **●ROC Area**
- **最も低いJRip (0.93)**
- **最も高いRandomForest (0.99).**
- excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), poor (0.60-0.70), fail (0.50–0.60).

- RQ3

同じデータセット、特性、上記の実験から得られたパフォーマンスが最も高い分類器を使用して句動詞の予測結果は何ですか？

The distribution of phrasal verbs' collocations

headword	EVP	A1 maxf	A1 type	A1 tfre	A2 maxf	A2 type	A2 tfre	B1 maxf	B1 type	B1 tfre	B2 maxf	B2 type	B2 fre
come from	A1	195	423	3022	203	384	2204	34	525	1558	29	366	819
look after	A2	314	269	1837	70	190	485	20	225	427	32	145	266
carry out	B1	299	186	2072	191	143	1560	183	186	1741	55	115	691
stand for	B2	105	91	370	10	69	103	59	88	204	33	39	80



6. 結論

- Nation（2013）とHawkins（2009）のモデルに従った、動詞のコロケーション使用は、動詞と句動詞のレベルを識別するための有効な基準特性であることを証明した。
- 動詞に共起する名詞のCFERれべう、頻度、タイプ数は分類に役たつ。
- 動詞および句動詞のCEFR-Jレベルを予測Wekaアルゴリズムは、TP Rate、Recall、Presion、F-measureおよびROC Areaに基づいて、ランダムフォレストアルゴリズムはより良い分類精度と予測パフォーマンスを示した。

- しかし、分類性能を改善する機会がある。まず、動詞が自動詞か自動かはこの研究では考慮されていない。また、この調査ではカテゴリーや機能にかかわらず、動詞のCEFR-Jレベルのみを使用した。将来、機械分類の精度を向上させるために、動詞のより質的な特性を考慮に入れることができる。
- 上記のいくつかの制限があるが、この研究の結果は動詞と句動詞のレベルを区別するためのより信頼性の高い客観的な方法を提供することが期待される

参考文献

- Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. (pp.i-v). Amsterdam: John Benjamins.
- Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1), 23-35.
- Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb– noun collocations. *Language Teaching Research*, 18(1), 54-74.
- Celce-Murcia, M., & Rosensweig, F. (1989). Teaching vocabulary in the ESL classroom. In M. Celce-Murcia & L. McIntosh (Eds.), *Teaching English as a second or foreign language* (pp. 241-257). New York: Newbury House Publishers Inc.
- Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M., & Papo, E. (1987). *Un niveau seuil. Paris*: Didier.

ご清聴ありがとうございます