11. パラレルコーパスを利用した検索と対数尤度比検定による対訳抽出

内山将夫@NICT
mutiyama@nict.go.jp

パラレルコーパスの英語学習への利用

- 辞書を調べるような感覚でパラレルコーパスを調べる
- 自由に利用できる対訳コーパス検索システム
  「日英対応付けコーパスの検索」

  http://www.kotonoba.net/~snj/cgi-bin/text-search/text-search.cgi

日英対応付けコーパスの検索サイト

- 35万文の対訳コーパス
  - 読売新聞とThe Daily Yomiuri 18万文
  - ロイター日英 7万文
  - 小説等 10万文
- 月間で200〜300人程度が使っている模様

検索される語の例

- 「バランス.*取」「Maintaining a balance between the technocrats' liberalism and economic nationalism, and curbing corruption among the nationalists was of keyimportance to the stability and economic development of Indonesian society.」「こうした構図にあっては、テクノクラートの自由主義と経済ナショナリズムの＜＜＜バランスがうまく取＞＞＞れ、しかもナショナリストに伴いがちな腐敗・汚職がそこそこコントロールされていることが、社会の安定と経済の発展には重要であった。」
- 「んじゃない」「you must not touch her.」「さわる＜＜＜んじゃない＞＞＞。」「But, I think the public might be convinced if everyone nominated Mr. Nakasone for the premiership.」「本当は中曽根さんを皆で担いだら、国民も納得する＜＜＜んじゃない＞＞＞か。」

英語教育への利用

- 日本大学中條清美先生

  http://www5d.biglobe.ne.jp/~chujo/resorce.html

  - パラレルコーパスを利用した語彙指導タスク集
  - パラレルコーパスを利用した文法指導タスク集1
  - パラレルコーパスを利用した文法指導タスク集2

タスクの例

1. 「decline」の日本語訳で多いものをあげてみよう.「下落」「減少」「衰退」「低下」「落ち込み」
2. 「efficiency」の日本語訳で多いものをあげてみよう.「効率」「燃費」「能率」
3. 「製品」にあたる英語で特に多いものは何ですか.「product」「products」
4. どんな製品がありますか.「(...) products」となるものを3つ見つけて，日本語訳もつけましょう.「foreign products (外国製品)」「industrial products (工業製品)」「oil products (石油製品)」「steel products (鉄鋼製品)」
5. 「commercial (...)」という用例を2つ見つけて日本語訳をつけよう.「commercial areas (商業地)」「commercial bank (都市銀行)」
6. 「inventory (...)」という用例を2つ見つけて日本語訳をつけよう.「inventory adjustment (在庫調整)」「inventory index (在庫指数)」
7. 「(...) access」という用例を2つ見つけて日本語訳をつけよう.「free access (自由なアクセス)」「Internet access (インターネットアクセス)」

対訳候補の抽出

対訳コーパスを便利に使うには,
「efficiency」の日本語訳で多いものをあげてみよう
という質問に対して,
「効率」「燃費」「能率」
という回答が，素早く分かることが必要である.
そのために,

1. decline と特に良く共起する日本語単語を抽出する
2. その共起の度合を表す尺度として対数尤度比を利用する

対数尤度比を利用した対訳候補の抽出法

	効率	効率以外	
efficiency	a	b	a+b
efficiency 以外	c	d	c + d
	a+c	b+d	n

a = 「効率」と「efficiency」が共に存在する対訳文の数
b = 「効率」がなく「efficiency」がある対訳文数
c = 「効率」があり「efficiency」がない対訳文数
d = 「効率」も「efficiency」も存在しない対訳文数
n = 全対訳文数

読売新聞と The Daily Yomiuri のデータでは

$$a = 137,\ b = 59,\ c = 284,\ d = 149520$$

もし,「効率」の存在が「efficiency」の存在に影響を与えないならば

$$\frac{a}{a+c} \sim \frac{b}{b+d} \tag{1}$$

のはずである．しかし，もし,「効率」と「efficiency」が良く共起するなら(あるいはあまり共起しないなら)

$$\frac{a}{a+c} \neq \frac{b}{b+d} \tag{2}$$

のはずである．
1 式は両者が確率的に独立であり，2 式は両者が確率的に従属であることを示す．→ 独立性の検定を利用し，独立性が低いものを抽出する．

対数尤度比検定 (Log-Likelihood Ratio Test)

$$LLR = \log \frac{P(\text{データ}\,|\,\text{従属})}{P(\text{データ}\,|\,\text{独立})} \tag{3}$$

を計算する．$LLR \gg 0$ならば,「効率」と「efficiency」については，従属であると考えた方が良いので，この値が大きければ，対訳候補として有望と考える．つまり各単語と「efficiency」について，LLR を計算し，その LLR が大きい単語を対訳候補とする．

対数尤度比の例

単語	対訳	a	b	c	d	LLR
efficiency	効率	137	59	284	149520	710
efficiency	化	79	117	6984	142820	115
efficiency	燃費	14	182	15	149789	73
efficiency	性	51	145	4861	144943	67
efficiency	向上	22	174	455	149349	58
decline	減少	139	487	525	148849	439
decline	低下	91	535	550	148824	245
decline	下落	81	545	414	148960	230
decline	減	56	570	245	149129	165
decline	連続	57	569	514	148860	131
commercial	商業	120	325	101	149454	564
commercial	銀行	131	314	1900	147655	302
commercial	都市	50	395	710	148845	111
commercial	捕鯨	26	419	83	149472	92
commercial	都銀	23	422	44	149511	91

データの表現法

P(データ | 独立) や P(データ | 従属) を計算するには，データを数値表現しないといけない．このときのデータの単位は対訳文である．そこで

$$E_i = \begin{cases} 1 \text{「efficiency」が対訳文} i \text{に出現する} \\ 0 \text{ 出現しない} \end{cases}$$

$$K_i = \begin{cases} 1 \text{「効率」が対訳文} i \text{に出現する} \\ 0 \text{ 出現しない} \end{cases}$$

という変数を定義する．すると

$$a = \sum_{i=1}^{n} [E_i = 1][K_i = 1] \tag{4}$$

$$b = \sum_{i=1}^{n} [E_i = 1][K_i = 0] \tag{5}$$

$$c = \sum_{i=1}^{n} [E_i = 0][K_i = 1] \tag{6}$$

$$d = \sum_{i=1}^{n} [E_i = 0][K_i = 0] \tag{7}$$

である．

$P(\text{データ}\,|\,\text{独立})$の計算

$$
\begin{aligned}
\log P(\text{データ}\,|\,\text{独立}) &= \sum_{i=1}^{n} \log P(E_i, K_i|\,\text{独立}) \\
&= a \log P(E_i = 1, K_i = 1|\,\text{独立}) \\
&+ b \log P(E_i = 1, K_i = 0|\,\text{独立}) \\
&+ c \log P(E_i = 0, K_i = 1|\,\text{独立}) \\
&+ d \log P(E_i = 0, K_i = 0|\,\text{独立})
\end{aligned}
$$

$$
\begin{aligned}
&\log P(E_i = 1, K_i = 1|\,\text{独立}) \\
&= \log P(E_i = 1|\,\text{独立}) P(K_i = 1|\,\text{独立}) \\
&= \log \frac{a+b}{n} \frac{a+c}{n}
\end{aligned}
$$

$$
\begin{aligned}
&\log P(E_i = 1, K_i = 0|\,\text{独立}) \\
&= \log P(E_i = 1|\,\text{独立}) P(K_i = 0|\,\text{独立}) \\
&= \log \frac{a+b}{n} \frac{b+d}{n}
\end{aligned}
$$

残りの2つについても同様

$P(\text{データ}\,|\,\text{従属})$の計算

$$
\begin{aligned}
\log P(\text{データ}\,|\,\text{従属}) &= \sum_{i=1}^{n} \log P(E_i, K_i|\,\text{従属}) \\
&= a \log P(E_i = 1, K_i = 1|\,\text{従属}) \\
&+ b \log P(E_i = 1, K_i = 0|\,\text{従属}) \\
&+ c \log P(E_i = 0, K_i = 1|\,\text{従属}) \\
&+ d \log P(E_i = 0, K_i = 0|\,\text{従属})
\end{aligned}
$$

$$\log P(E_i = 1, K_i = 1|\,\text{従属}) = \log \frac{a}{n}$$

$$\log P(E_i = 1, K_i = 0|\,\text{従属}) = \log \frac{b}{n}$$

残りの2つについても同様

問題 (15分)

$$LLR = \log \frac{P(\text{データ} \,|\, \text{従属})}{P(\text{データ} \,|\, \text{独立})}$$

を a, b, c, d, n により，なるべく簡単な形式で表現して下さい.

回答例

$$
\begin{aligned}
LLR &= a\log\frac{\frac{a}{n}}{\frac{a+b}{n}\frac{a+c}{n}} + b\log\frac{\frac{b}{n}}{\frac{a+b}{n}\frac{b+d}{n}} \\
&+ c\log\frac{\frac{c}{n}}{\frac{c+d}{n}\frac{a+c}{n}} + d\log\frac{\frac{d}{n}}{\frac{c+d}{n}\frac{b+d}{n}} \\
&= a\log a + b\log b + c\log c + d\log d \\
&+ (a+b+c+d)\log n \\
&- (a+b)\log(a+b) - (a+c)\log(a+c) \\
&- (b+d)\log(b+d) - (c+d)\log(c+d) \\
&= l(a) + l(b) + l(c) + l(d) + l(n) \\
&- l(a+b) - l(a+c) - l(b+d) - l(c+d)
\end{aligned}
$$

ただし，$l(x) = x\log(x)$

まとめ

- 対訳コーパスは，英語教育や日本語教育にも役立つ
- LLRを利用することにより，対訳候補を抽出できる