共起語の関係強度の分析

5K－3

宮原　豊，寺本　陽彦，松本　俊二
Yutaka MIYAHARA, Youhiko TERAMOTO, Shunji MATSUMOTO
富士通㈱計算科学技術センター

1.　はじめに

自然言語処理で、同一文書中に出現する二語（共起語）の出現頻度に基づいて語の意味的関係を推測する共起語分析がある。しかし、どんな語とも結びつく一般的な語も共起頻度は高いため、単なる頻度情報だけでは意味的関係のある共起語を特定することはできない。これに対して、単語の文書集合での頻度分布の情報を利用して単語間の関連性を定量化する手法がある。本稿では、共起語の独立性に着目して共起語の意味的な関連性を分析する手法について報告する。

2.　共起語の独立性

文書集合中の共起語の出現頻度を、各語出現／非出現をマトリックスにした分割表に整理すると、語の出現パターンに応じて三種類の独立性があることがわかる。

【表１：独立性タイプ１】

t1＼t2	出現	非出現	t1 計
出現	10	100	110
非出現	100	1000	1100

（タイプ１）表１では、t1 出現文書 110 件中の t2 出現／非出現文書比率は 1:10 で、t1 非出現文書 1100 件中の同比率も 1:10 である。t1 出現文書でも t1 非出現文書でも t2 の出現確率は一定であり、t1 と t2 の独立性は高い。

【表２：独立性タイプ２】

t3＼t4	出現	非出現	t3 計
出現	10	10	20
非出現	100	1000	1100

（タイプ２）表２では、t3 出現文書 20 件中の t4 出現／非出現文書比率は 1:1 であるのに対し、t3 非出現文書 1100 件中の同比率は 1:10 である。t3 出現文書では t3 非出現文書と較べて t4 の出現確率が高く、t3 と t4 は正の相関にあり独立性は低い。

【表３：独立性タイプ３】

t5＼t6	出現	非出現	t5 計
出現	1	100	101
非出現	100	1000	1100

（タイプ３）表３では、t5 出現文書 101 件中の t6 出現／非出現文書比率は 1:100 であるのに対し、t5 非出現文書 1100 件中の同比率は 1:10 である。t5 出現文書では t5 非出現文書と較べて t6 の出現確率が低く、t5 と t6 は負の相関にあり独立性は低い。

そこで、独立性の高いタイプ１や負の相関にあるタイプ３の共起語は意味的関係が弱く、独立性が低く正の相関にあるタイプ３の共起語は意味的関係が強いという仮定にたって関係強度を分析する。

3.　独立性に基づく共起語の関係強度

独立性を定量的に評価するために統計的検定を行う。ここでは確率分布の適合度や独立性の検定に一般に利用される χ^2 検定を使う。検定により独立性仮説が棄却された中で正の相関にあるものが関係強度の強い共起語である。以下、共起語の出現頻度を分割表に整理して χ^2 値を求め仮説検定する。さらに仮説の棄却域を利用した関係強度の指標を導入する。

(1) 分割表に整理

分析対象の共起語の出現文書件数を表４の分割表に整理する。f_{ij} は分割表上の各度数、f_i、f_j は i、j を固定した場合の周辺度数、n は全度数である。

【表４：分割表】

i＼j	1	2	i 周辺度数
1	f_{11}	f_{12}	f_1
2	f_{21}	f_{22}	f_2
j 周辺度数	f_1	f_2	n

(2) χ^2 値の計算

式１で分割表の χ^2 値を計算する。

$$\chi^2 = \sum_i \sum_j \frac{(n f_{ij} - f_i f_j)^2}{n f_i f_j} \quad \cdots\cdots(式1)$$

A method for analyzing relationship between cooccurrent words.
Yutaka Miyahara, Youhiko Teramoto, Shunji Matsumoto
Computational Science and Engineering Center, Fujitsu Ltd.
9-3, Nakase 1, Mihama-ku, Chiba, 261-8588, Japan

(3) 仮説検定

χ^2分布表を利用して自由度1のχ^2分布と比較する。

【表5：χ^2分布表（自由度1）】

α	0.500	0.400	0.300	0.200	0.100
χ^2	0.4549	0.7083	1.074	1.642	2.706
α	0.050	0.025	0.010	0.005	0.001
χ^2	3.842	5.024	6.635	7.879	10.83

χ^2分布表は上側確率αに対応するχ^2値（パーセント点）を与える統計表で、2×2 分割表の場合自由度は1である。χ^2分布表を参照してχ^2値があるαでパーセント点を越える場合は有意水準αで独立性の仮説を棄却できる。χ^2値が2.8の時、αが0.100、0.050 のパーセント点はそれぞれ 2.706、3.842 であるため、有意水準 10%では棄却できるが 5%ではできない。有意水準は正しい仮説を誤って棄却する危険率に相当するため、危険率 10%（10 に 1 つの誤りを認める）では棄却できるが、5%（20 に 1 つの誤りを認める）では棄却できないことになる。

(4) 仮説の棄却域を利用した関係強度の指標

有意水準を動的に変えることにより仮説の棄却域が変化することを利用して、「独立性の仮説が棄却可能な有意水準の最小値」を共起語の関係強度の指標として導入する。この値はパーセント点がχ^2値と等しくなる上側確率の値である。指標値が小さいほど仮説を誤って棄却する危険率が低いため独立性が低い、つまり関係強度が強いとみなすことができる。

4. 実験と考察

計算機のサポートに関する質問応答事例 12,000 件に基き、ある RDBMS 製品名「製品 A」の共起語を分析した。共起数が 6 以上の共起語についてχ^2値と上側確率、相関の正負を調べ、表6、7に共起数 20 未満（104 語）と 20 以上（35 語）に分けて上側確率の昇順に等間隔で 10 件ずつ抽出した。

【表6：共起数20未満の共起語】

共起語	χ^2値	上側確率	共起数	相関
ｲﾝｽﾀﾝｽ	113.5	1.69e-24	10	＋
表領域	55.97	7.36e-12	9	＋
CONNECT	26.49	2.65e-05	11	＋
ﾌｧｲﾙ	21.94	2.81e-4	8	－
ｴﾗｰｺｰﾄﾞ	5.117	2.369	9	＋
発行	2.377	12.31	6	－
正常	0.8146	36.68	17	－
ﾊﾞｯｸｱｯﾌﾟ	0.6641	41.51	6	＋
OS	0.3131	57.58	17	＋
機種	0.05287	81.81	8	－

【表7：共起数20以上の共起語】

共起語	χ^2値	上側確率	共起数	相関
ORA	1361	4.9e-296	199	＋
00600	171.8	2.95e-37	20	＋
SQL	71.14	3.33e-15	54	＋
ODBC	68.84	1.07e-14	26	＋
ﾃﾞｰﾀ	31.52	1.98e-6	54	＋
表示	13.93	0.01898	28	－
ｴﾗｰ	5.312	2.118	120	＋
使用	2.517	11.26	138	＋
ﾌﾟﾛｸﾞﾗﾑ	0.7582	38.39	20	＋
原因	0.02593	87.21	20	－

共起数が違う表6、7とも、上側確率が低く正の相関の語はエラーコードやDB用語など「製品A」と意味的に関係ある語が多く、上側確率が高い語は質問応答に一般的に現れる語が多い。この指標は共起数に関わらず、共起語の意味的な関係強度の実態を全般的に反映していることがわかる。

しかし、「データ」や「エラー」のように上側確率が低くても特に意味的な関係がないと思われる語もある。これは、これらの語が文書集合で「製品 A」と随伴する傾向が強いことを示しており、表現の定型性（「製品 A でエラー XXXXX が発生」）など文書集合の特性の影響を無視できないことがわかる。

5. まとめ

独立性が低く正の相関にある共起語は意味的関係があるという仮定に基づく本手法により、共起数という表面的な頻度情報ではわからない共起語の意味的な関係を抽出できた。文書集合の特性により一般的な語の関係強度が不適切に強くなる場合があるが、これは統計情報から意味的関係を推測しようとする手法に本質的に付随する困難である。

今回導入した指標は、意味的な関係の強弱を相対的に知る目安となるが、指標値がいくらならどのくらい関係が深いという絶対的な基準は明らかでない。実用化を考慮して、人間の判断に近い絶対的な関係強度の評価手法を検討することが今後の課題である。

参考文献：藤江、兵藤、池田：文字、単語統計解析の一手法、第51回情処全国大会、1H-6、1995