自然言語処理 113－6
音声言語情報処理 11－15
（1996. 5. 28）

ニューラルネットワークを用いたコーパスからの共起強度の抽出

中川 賢一郎, 山口 昌也, 乾 伸雄, 野瀬 隆, 小谷 善行, 西村 恕彦
(東京農工大学 工学部 電子情報工学科)

かな漢字変換の同音語における曖昧性を解消するために、共起ネットワークを用いる手法を提案する。共起ネットワークは、分類語彙表に対応した単語の分類番号をノード、共起強度を有向アークとした相互結合型ニューラルネットであり、コーパスから比較的容易に抽出することができる。この共起ネットワークを実際にコーパスから作成し、未知の文章に対して次に続くべき単語の分類番号を予想させる実験を行うと、10位までの候補に12%の割合で正解を入れることができた。また、かな漢字変換の同音語処理に応用した結果、未知の文章に対して71%の割合で正解を選択することができた。

Extraction of Strength of Cooccurrence from Corpora using a Neural Network

Kenichiro NAKAGAWA, Masaya YAMAGUCHI, Nobuo INUI, Takashi NOSE,
Yoshiyuki KOTANI, Hirohiko NISIMURA
(Dept. of Computer Science, Tokyo University of Agric. and Tech.)

In order to disambiguate homography words in Kana-Kanji translation, we propose a technique using a *concurrent network*. A *concurrent network* is a mutually-connected neural network with nodes of classification number of the word and directed arcs which means strength of cooccurrence. It is easy to extract the strength from corpora. We made the network from corpora, and we made the experiment to guess the classification number of the word which should come to next in the unknown sentences. The experiment shows that the system can put a correct answer up to the 10th candidates at the rate of 12%. Moreover, we made the experiment to apply the network to disambiguate homography words in Kana-Kanji translation. In this case, the system can choose the correct answer at the rate of 71% against the unknown sentences.

1 まえがき

べた書きかな漢字変換などを複雑にする要因として、同音語による曖昧性や分かち書き処理における曖昧性がある。この同音語による曖昧性を除去するために、その単語の位置や頻度を用いる方法、単語の持つ共起関係を利用する試みがなされている [1][2][4]。また、コネクショニストモデルやニューラルネットワークといった神経生理学的特徴を持つ知識構造を利用したものもある [10][11][12][13]。これらは、同音語の曖昧性除去だけではなく未知語処理の分野でも用いられている [8]。

本稿では、共起関係を共起ネットワークと呼ぶ知識表現によって表す。共起ネットワークとは、脳の連想行為をモデル化するために設計されたネットワークで

あり、単語の持つ分類番号をノード、共起強度を有向アークと見なした知識構造である。このネットワークをニューラルネットワークと見なし、カオスニューラルネットワーク [9] の想起のアルゴリズムを用いることでコーパスから共起強度を抽出することができる。この共起ネットワークの特徴は、あるノードから伸びるアークのアーク重み (共起強度) の総計が 1 となることと、アークが抑制リンクの形をとらないことである。このため、共起ネットワークは一種の Bayesian ネットワークととらえることができ、ノードから不確定的にパルスが伝達される場合に有益な知識表現である。

本稿では、この共起ネットワークをコーパスから抽出する手法を提案し、この知識表現を相互結合型ニューラルネットワークと見なすことにより行った連想実験の結果と、かな漢字変換の同音語処理に応用した場合の効果を報告する。

2 共起ネットワーク

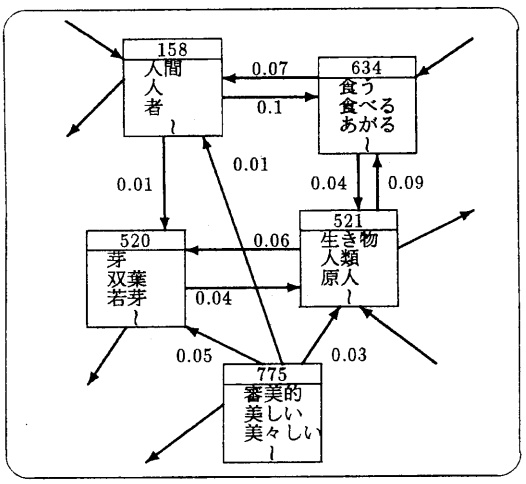

図 1 共起ネットワークの一部

共起ネットワークとは、単語の持つ分類番号をノード、共起関係を表す有向アークによって構成されている (図 1)。図のアークが持つ数値は、そのアークに含まれる単語が共通に持つ分類番号である。分類番号は分類語彙表 [14] の分類番号 (合計 822 個) と一対一対応しているが、シソーラス上の階層構造は考慮していない。各アークは共起強度を表すアーク重みを持っている。共起強度は 0 から 1 までの数値をとる値であり、共起が強いほど大きな値をとる。また、一つのノード (分類番号) から伸びるアークの共起強度の総計は 1 となるように設定されている。

共起ネットワークの持つ特徴として、

- 意味的に近い単語を一つの分類として扱っているために、それほど大規模なネットワークを必要としない。
- 複数単語から共起する単語 (話しの流れから共起する単語) を見つけることが容易である。

という利点がある。

3 共起ネットワーク作成方法

3.1 カオスニューラルネットワーク

共起ネットワークは次に示すカオスニューラルネットワークのモデルを用いている。

$x_i(t+1)$	:	i 番目のノードの時刻 (t+1) における出力値
n	:	ネットワークを形成するノードの総数
w_{ij}	:	j 番目から i 番目ノードへの重み
α	:	不応性項をスケーリングするパラメータ
k	:	時間的減衰定数
ϵ	:	シグモイド関数の鋭さを表す変数
θ	:	ノードしきい値

$$x_i(t+1) = f(\sum_{j=1}^{n} w_{ij} \sum_{d=0}^{t} k^d x_j(t-d) - \alpha \sum_{d=0}^{t} k^d x_i(t-d) - \theta) \quad (1)$$

$$= f(I_i(t+1) + J_i(t+1) - \theta) \quad (2)$$

$$f(y) = \frac{1}{1+exp(\frac{-y}{\epsilon})} \quad (3)$$

$I_i(t), J_i(t)$ は時刻 t にノード i が持つ内部状態変数値であり、そのノードの持つ興奮値と考えることができる。特に J は不応性を表す変数であり、この値がカオス的な振る舞いを引き起こす。

I, J の式は次のようになる。

$$I_i(t+1) = \sum_{j=1}^{n} w_{ij} \sum_{d=0}^{t} k^d x_j(t-d) \quad (4)$$

$$= kI_i(t) + \sum_{j=1}^{n} w_{ij} x_j(t) \quad (5)$$

$$J_i(t+1) = -\alpha \sum_{d=0}^{t} k^d x_i(t-d) \quad (6)$$

$$= kJ_i(t) - \alpha x_i(t) \quad (7)$$

ノードの出力値 x は単位ステップ関数ではなく、0から1までの連続した数値をとることが他のニューロンモデルとの大きな違いである。

3.2 共起ネットワーク作成アルゴリズム

ここではコーパスから共起ネットワークを作成するアルゴリズムを示す。

[準備] 入力するテキストコーパスを形態素解析し、自立語以外を削除し、対応する分類番号を付加する。

初期化としてすべてのノード i において

$$I_i(t) = J_i(t) = x_i(t) = out\_x_i = 0$$

$$I_i(t+1) = J_i(t+1) = x_i(t+1) = 0$$

を行う。$I(t+1)$, $J(t+1)$, $x(t+1)$ は一時刻先の $I(t)$, $J(t)$, $x(t)$ の計算に用いる変数である。また、すべてのノード i, j において

$$w\_tmp_{ij} = 0$$

を行う。w_tmp, out_x はアーク重みである w の計算で用いる変数である。

[手順 1] テキストコーパスから一単語取り出す。その単語の分類番号を b とする。

[手順 2] すべてのノード i において

$$I_i(t+1) = kI_i(t) + \sum_{j=1}^{n} w_{ij} x_j(t)$$

$$J_i(t+1) = kJ_i(t) - \alpha x_i(t)$$

$$x_i(t+1) = f(I_i(t+1) + J_i(t+1) - \theta)$$

を行う。次に、すべてのノード i において

$$I_i(t) = I_i(t+1)$$

$$J_i(t) = J_i(t+1)$$

$$x_i(t) = x_i(t+1)$$

を行う。

[手順 3] すべてのノード i において

$$out\_x_i = out\_x_i + x_i(t)$$

$$w\_tmp_{bi} = w\_tmp_{bi} + x_i(t)$$

を行う。すべてのノード i と j において

$$w_{ij} = \frac{w\_tmp_{ij}}{out\_x_j}$$

を行う。

[手順 4]

$$I_b(t) = I_b(t) + 1$$

を行う。

[手順 5] 手順1から手順4までの作業をテキストコーパスの最後まで行う。テキストコーパスが終わった時点で終了し、w_{ij} のデータをファイル出力する。

手順 2 は入力された単語から与えられた刺激が、ネットワーク中を伝搬していくことを表している。手順3は興奮しているノードほど、入力された単語が属するノード b との結びつきが強くなることを表している。図 2 は上の学習のプロセスを概念化したものである。

入力された単語が二重丸の分類ノードに含まれているとする (①)。このネットワークの各ノードの内部状態変数値 (興奮度) を用いて、一時刻分だけ想起させると各ノードの内部状態変数値は変化し、②の状態になったとする (ノードの数値は新しく更新された内部状態変数値 I と J の和)。各ノードの内部状態変数値に見合った分、二重丸のノードへのつながりを強くする (③)。最後に、二重丸のノードの内部状態変数値に 1 を加える (④)。

これらの作業を繰り返すことにより、コーパスから共起関係を共起ネットワークという形で獲得することができる。

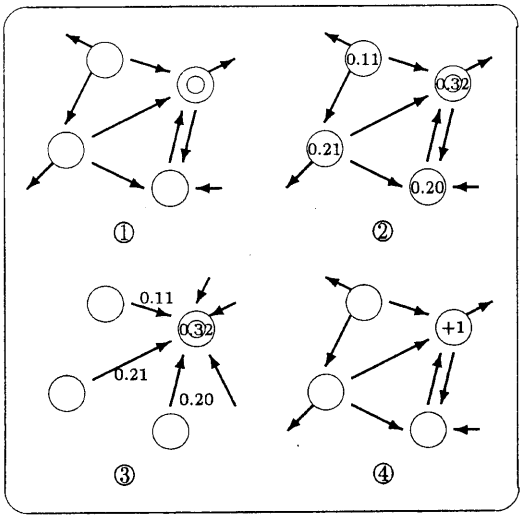

図 2 学習の概念図

4 共起ネットワークを用いた想起

本節では、前項で作成した共起ネットワークを用いて、連想実験やかな漢字変換実験を行うためのアルゴリズムを示す。

[準備] 入力するテキストを形態素解析し、自立語以外を削除し、対応する分類番号を付加する。

初期化としてすべてのノード i において

$$I_i(t) = J_i(t) = x_i(t) = 0$$

$$I_i(t+1) = J_i(t+1) = x_i(t+1) = 0$$

を行う。

また、ファイルからすべての i, j における w_{ij} (共起ネットワークの全アーク重みに相当) を読み込んでおく。

[手順 1] 前項の [手順 1] と同様。

[手順 2] 前項の [手順 4] と同様。

[手順 3] 前項の [手順 2] と同様。

[手順 4] 手順 1 から手順 3 までの作業をある時点まで行い、その時点でのネットワーク中のどこが興奮しているかを調べることで、次に来る単語の分類を予想することができる。

5 評価実験

5.1 共起ネットワークの抽出

文献 [16] の文章をコーパスとして用い、共起ネットワークを抽出した。なお、本システムは NWS-5000 上の C 言語を用いて実現されている。システムの入力と出力を表 1 に示す。

表 1 入力データと出力データ

入力	コーパス名	「子供と自然」[16]
	文数	2927 文
	単語数	55518 単語
出力	ノード数	822 個
	アーク数	674862 本

各パラメータの値は次のものを用いた。

$$\alpha = 0.4, \quad k = 0.7, \quad \epsilon = 0.05, \quad \theta = 0.5$$

5.2 共起ネットワークを用いた連想実験

前項で作成した共起ネットワークを用い、システムに文章を読ませていくことで、次に来る単語の分類を予想させる実験を行った。予想の順位は、そのときの共起ネットワークのノード内部状態変数値の大きい順である。

実験は、

環境 1 ... コーパスの全文を用いて作成した共起ネットワークに、入力テキストとして同じコーパスの全文を用いる。

環境 2 ... コーパスの前半部文を用いて作成した共起ネットワークに、入力テキストとして同じコーパスの後半部分を用いる。

環境 3 ... コーパスの全文を用いて作成した共起ネットワークに、入力テキストとして朝日新聞の社説を用いる。

の三つの環境で行った。それぞれの場合の、予想の順位における正解の分布を図3〜図5で示す。

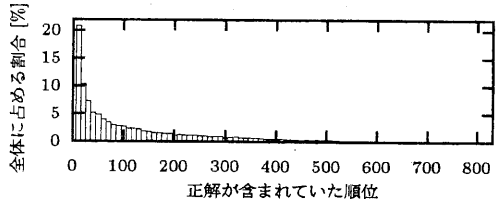

図3 環境1での正解の分布

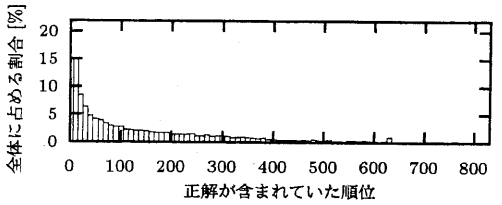

図4 環境2での正解の分布

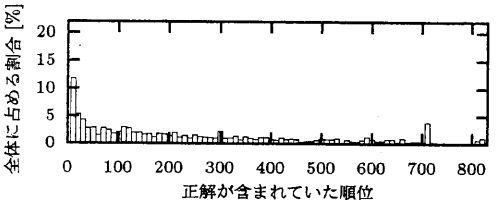

図5 環境3での正解の分布

この実験とは別に、文章を構成する自立語列を示し、次に続くと思われる自立語を1位から10位まで自由に書いてもらう実験を人間に対し行った。10位までの間に、実際に続く自立語と同じ分類番号を持つ自立語が入っていれば正解とし、6人に対しそれぞれ20問ずつ出題した。その結果、人間は予想順位の10位までに31%の割合で次に来るべき分類の単語を入れることができた。

予想の1位から10位までの間に正解が含まれている割合を比較したものが表2である。

表2 連想実験の結果

	1〜10位に正解の分類が入っていた割合
人間	31%
環境1	21%
環境2	15%
環境3	12%

5.3 かな漢字変換における同音語候補の選択実験

共起ネットワークを用いて、かな漢字変換における同音語の候補の中から最も適切な(内部状態変数値が最も高い分類に含まれる)候補を選択する実験を行った(図6)。また、bi-gram共起確率を共起ネットワークの学習に用いたコーパスから抽出し、その確率を用いて同音語を選択させる実験も行った(図7)。共起ネットワーク、bi-gram共起確率を用いた実験共に、前出の三つの環境下で行った。

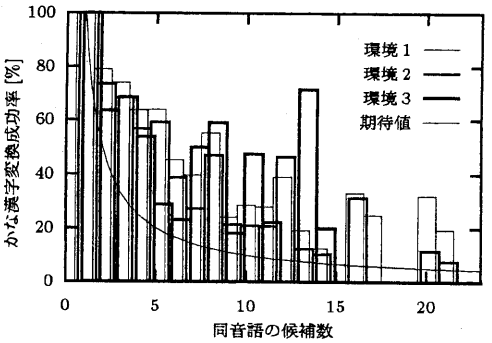

図6 共起ネットワークを用いた同音語処理における候補数に対する成功率の変化

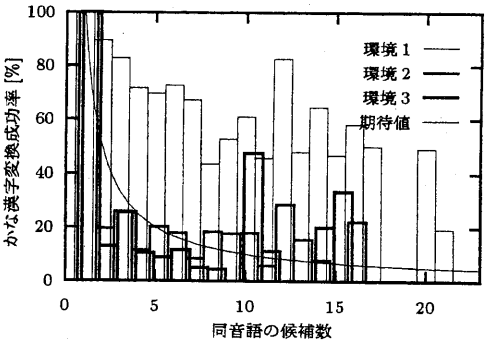

図7 bi-gramの共起確率を用いた同音語処理における候補数に対する成功率の変化

bi-gramの共起知識を用いた同音語選択では、環境1での正解率は高いが、環境2,3での正解率は極端に下がり、共起ネットワークを用いた同音語選択では、環境2,3でも極端には正解率は下がらないことが示された。

入力テキスト全体での同音語選択の成功率は、表3で示す。

表3 かな漢字変換に成功した割合

	bi-gram	共起ネットワーク
環境1	88%	81%
環境2	55%	78%
環境3	47%	71%

6 むすび

本稿では、共起知識を表現するツールとして共起ネットワークを提案し、これをコーパスから抽出する手法を示した。この共起ネットワークを用いて想起をしながら文章を読み進めることによって、次に続く単語の分類番号を10の予想候補中に12%の割合で入れることができた。さらにこの知識表現をかな漢字変換の同音語選択に用いた結果、71%の割合で正しい候補を選択することができた。

謝辞

本研究にあたり、適切な助言と実験を手伝って下さった東京農工大学工学部電子情報工学科西村・小谷研究ユニットの皆様に感謝します。

参考文献

[1] 本間 茂, 山階 正樹, 小橋 史彦：連語解析を用いたべた書きかな漢字変換, 情報処理学会論文誌, Vol 27, No 11, pp.1062–1067, (1986).

[2] 山本 喜大, 久保田 淳市：共起グループを用いたかな漢字変換, 情報処理学会第44回全国大会論文集, 4p–11, pp.189–190, (1992).

[3] 工藤 育男, 井ノ上 直己：コーパスに基づく共起知識の獲得とその応用, 人工知能学会誌, Vol 10, No 2, pp.205–211, (1995).

[4] 高橋 雅仁, 吉村 賢治, 首藤 公昭：共起情報を用いた同音語処理, 自然言語における文脈シンポジウム発表論文, (1995).

[5] 池原 悟, 白井 諭, 河岡 司：大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出法, 電子情報通信学会技術報告, NLC95–3, pp.17–24, (1995).

[6] Shiho Nobesawa : Segmenting a sentence into morphemes using statistic information between words, COLING 94 PROCEEDINGS, Vol 1, pp.227–233, (1994).

[7] 有田 英一, 岡 隆一：新聞記事データからの断片的知識の連鎖の抽出, 電子情報通信学会技術報告, NLC93–66, pp.23–30, (1993).

[8] 上田 一人, 瀧口 伸雄, 小谷 善行：日本文における共起情報を用いた未知語検索, 情報処理学会第46回全国大会論文集, 2b–6, pp.101–102, (1992).

[9] 渡辺 正峰, 合原 一幸, 渡辺 駿介：カオスニューラルネットワークによる自動学習, 電子情報通信学会論文誌, Vol J78–A, No 6, pp.686–691, (1995).

[10] 小嶋 秀樹, 古郡 廷治：テキストの曖昧性を知識と文脈によって解消する計算モデル, 情報処理学会論文誌, Vol 32, No 11, pp.1366–1373, (1991).

[11] 内海 彰, 堀 浩一, 大須賀 節雄：コネクショニストモデルによる文脈を考慮した自然言語インタフェース, 人工知能学会誌, Vol 7, No 5, pp.828–835, (1992).

[12] 田村 淳, 安西 祐一郎：Connectionist Model を用いた自然言語処理システム, 情報処理学会論文誌, Vol 28, No 2, pp.202–210, (1987).

[13] 鈴岡 節：コネクショニストモデルに基づく認識理解に適した連想メモリ, 情報処理学会論文誌, Vol 34, No 7, pp.1540–1548, (1993).

[14] 国立国語研究所：分類語彙表 [フロッピー版] 解説書, (1993).

[15] 橋本 三奈子, 桑畑 和佳子：計算機用日本語基本名詞辞書IPAL, 第13回技術発表会論文集, Vol 13, pp.65–76, (1994).

[16] 河合 政雄：子どもと自然, 岩波ジュニア新書.