コーパスから抽出した単語間類似度に基づく意味ネットワーク*

1H－7

廣 恵太　伊藤 毅志　古郡 廷治†

電気通信大学情報工学科‡

1 はじめに

本稿では，語彙間の連想関係を記述する方法として，コーパスから抽出した単語間類似度に基づく意味ネットワークの手法を提案する．また，この語彙体系の妥当性の検証として，構成した意味ネットワーク上の活性伝播から文脈情報の抽出を行う．

大規模な意味ネットワークを構成するとき，ネットワークのノード間の関係を客観的方法によって抽出する必要がある．小嶋・古郡 [1] そこで小嶋らは，意味ネットワークを客観的・規則的に構成する方法として，英語辞書による手法を提案している．

辞書を用いた場合，単語間の一般的で安定した関係を抽出することができる反面，既存の辞書の範疇の語彙関係しか捉えられないという問題がある．自然言語テキストは，文章の属する分野，またはジャンルによって用いられる言葉の種類も用法も異なると考えられ，分野依存の語彙体系が求められる．辞書を用いた方法でこれを実現するには，幾つもの辞書が必要となり，実用的でない．一方，本手法は，あらかじめ分野にあわせた語彙体系を容易に構築することができ，上述の問題の解決法となる．

2 単語間の類似度

コーパスからの知識獲得として，単語の共起関係から単語間の類似度を求める手法が提案されている．[2][3] 本稿では意味ネットワーク上のノード間の関係を捉える目的から，Dagan et al[2] の問題点を解消した手法により単語間類似度を求める．

単語間類似度は次にあげる 2 つの仮説に基づき，独自に形態素解析を行ったLOB コーパス (総語数 約 100 万語) から抽出する．

(i) ある単語はその周辺に出現する単語によって意味づけられる．

(ii) 共起パターン (co-occurence patterns) もしくは分布 (distribution) が類似している単語はその程度に応じて意味的にも類似している．

2.1 相互情報量の計算

まずはじめに，仮説 (i) により相互情報量を求める．Church and Hanks[4] によれば，共起頻度 $f(w_1, w_2)$ が特に小さい時に相互情報量が不安定となる．よってここでは，$f(w_1, w_2) > 40$ となる単語を計算の対象としている．さらに中距離間で相互依存し合う単語間の関係を抽出するために，窓の大きさ $w = 41$ (約 2 文の大きさ) として次式より相互情報量を求める．

*Semantic network based on word similarity extracted from corpus

†Keita Hiro, Takeshi Ito, Teiji Furugori

‡Department of Computer Science, University of Electro Communications 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182, Japan

similar word	sim
medical	0.314829
patient	0.289601
health	0.280356
ill	0.279773
hospital	0.278425

表 1: The similar words of doctor

$$I(w_1, w_2) = log_2 \left(\frac{\mathrm{N}}{\mathrm{w}-1} \frac{\mathrm{f}(\mathrm{w}_1, \mathrm{w}_2)}{\mathrm{f}(\mathrm{w}_1)\mathrm{f}(\mathrm{w}_2)} \right) \quad (1)$$

2.2 単語間類似度の計算

単語間の類似度は仮説 (ii) の考えに基づき，ある 2 語の類似度を，それぞれの単語と任意の第 3 語との間の共起パターン — 相互情報量による分布 — の近似度から求める．

Dagan らは単語 w_1 ・ w_2 間の共起パターン類似度 $csim(w_1, w_2)$ を 2式 により定義している．$csim$ は，それぞれの語と $w\{w : w \in$ 全語彙$\}$ との双方向の相互情報量 $I(w, w_1)$ と $I(w, w_2)$, 及び $I(w_1, w)$ と $I(w_2, w)$ の間の 2 つの比率から求められる．

2式の中で，$I(w, w_1)$ と $I(w, w_2)$ のいずれか片方が値を持たない場合，$csim$ の値は小さくなる．つまり，2式 は，各単語が持つ相互情報量の項目数 $N(w_i)$ が大きな単語に対して有利に働く．さらに $N(w_i)$ は，コーパス中での出現頻度 $f(w_i)$ と相関関係にある．その結果として，この手法を用いた場合，任意の単語と高頻度語との間の類似度は高くなる傾向がある．この傾向は，$N(w_i)$ にばらつきが出る小規模コーパスを用いる場合，特に問題となる．我々は，この問題を改善した $sim(w_1, w_2)$ を次式により定義し，，単語間類似度を求める．

$$sim(w_1, w_2) = 1000 \cdot \frac{csim(w_1, w_2)}{N(w_1) + N(w_2)}$$

表 1 は LOB コーパスから単語間類似度を抽出した結果から，単語 doctor との類似度上位語を示したものである．

3 単語間類似度に基づく意味ネットワーク

意味ネットワークは制限語彙集合 (SLDV) をノード，単語間類似度をリンクの重みとして構成される．制限語彙集合はコーパスベースト手法の本質的な問題である低頻度語の信頼性問題 (sparse data problem) を避け，頑健な語彙体系を構成するという目的により定義される．

$$sim(w_1, w_2) = \frac{\sum_w \min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w))}{\sum_w \max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w))} \quad (2)$$

3.1 制限語彙 SLDV

英語辞書 LDOCE[5] の意味カテゴリーにおける制限語彙集合である Longman Defining Vocabulary (LDV) から SLDV (Subset of LDV) を求める. SLDV は LDV 2,851 語のうち以下にあげる (1) 同形語の区別を行わない, (2) 低頻度語を, 派生関係にある上位頻度語に融合する, (3) 機能語を取り除く, という処理を行った 1,691 語の語彙集合である.

3.2 意味ネットワークの構成

構成するネットワークは SLDV 1,691 のノードと, ノード間を結ぶ 827,812 (約 490 / 1 ノード) のリンクからなる. ノード w_i と w_j を結ぶリンクは重み k_{ij} を持つ. k_{ij} は $sim_(w_i, w_j)$ と等しい. また各ノードは活性を持つことができる.

4 文脈情報の抽出

テキストの文脈情報は構成した意味ネットワーク上の活性パターンから抽出される. 本稿において抽出する文脈情報とは, テキストの内容を象徴する単語の集合である.

4.1 ネットワークの活性伝播

テキストの入力は, 意味ネットワーク上の対応するノードに活性エネルギーを加えることによって実現する. テキストは時系列に従い各時刻に一語づつ, 活性エネルギーとして入力される.

入力語に対応するノードに活性化エネルギーを加えると, リンクを介して隣接ノードに活性が伝播することによって活性パターンがつくられる. あるノード w_i の時刻 $t+1$ における活性度 $a_i(t+1)$ を次の 2 つの式から計算する.

$$\begin{aligned} a_i(t+1) &= \alpha \cdot a_i(t) + L_i(t) + E_i(t) \\ L_i(t) &= \sum_j a_j(t) \cdot k_{ij}(t) \end{aligned}$$

$$\begin{cases} \alpha & : \text{ノード内部の活性の減衰係数} \\ a_i(t) & : \text{ノード } w_i \text{ の活性度} \\ L_i(t) & : \text{リンクを介して } w_i \text{ に与えられる活性} \\ E_i(t) & : \text{外部から } w_i \text{ に加えられる活性} \end{cases}$$

4.2 文脈ベクトル

ある時刻での活性パターン $P(t)$ は, 直後の活性状態 $P(t+1)$ に影響を及ぼす. すなわち活性パターンは, 時間的系列に従い, 線条的に発展していく. 我々は, 活性パターン間のこのような時間的・線条的関係を文脈と呼び, その変化の履歴から文脈情報を求める.

時刻 t におけるネットワークの活性パターン $P(t)$ を, 各ノードが持つ活性 $a_i(t)$ $(1 \leq i \leq 1691)$ を次元成分とする 1,691 次元ベクトルにより表す.

$$P(t) = (a_1(t), a_2(t), \ldots, a_i(t), \ldots, a_{1,691}(t))$$

ここで, 時刻 t から $t+l$ までの間の文脈情報を, 各時刻にそれぞれのノード w_i が持つ活性の平均 $\sum_{k=1}^{1,691} a_i(k)/l$ から求める. この活性平均を次元成分とする文脈ベクトル C_l^{t+l} は以下のように P から導出される.

$$C_l^{t+l} = \frac{P(t) + P(t+1) + \cdots + P(t+l)}{l}$$

4.3 文脈情報の抽出

あるテキストは, 語彙体系に入力された結果, 文脈ベクトルに変換される. こうして得られた文脈ベクトルの次元成分のうち, $\max_m(\sum_{k=1}^{1,691} a_m(k)/l)$ となるノード m は, テキストにおいて最も頻繁に, または最も強く連想される単語である. 我々は, このような単語が文脈の内容を象徴しているという仮説に基き文脈情報を抽出する. さらに, その結果をテキストの表題や心理実験の結果と比較することによって, ネットワークの妥当性を検討する.

5 おわりに

本稿ではまず, コーパスから客観的・体系的な手法により抽出した単語間類似度に基づいた意味ネットワークを構成した. また, 構成したネットワーク上の活性パターンの履歴から文脈ベクトルを構成し, 文脈情報を抽出する方法について述べた. 本手法は, 客観的知識に基づいた語彙体系を構築する上で, 基本的な枠組となる. 今後はさらに,

(1) より妥当な単語間類似度の検討

(2) 制限語彙の拡大

(3) より大規模なコーパスの利用

(4) EDR 共起辞書の利用

等を試みた上で, 分野依存性の検証, 文間の結束性の研究などに応用してゆく.

参考文献

[1] 小嶋 秀樹, 古郡 廷治 : 単語の意味的な類似度の計算. 電子情報通信学会技術研究報告, AI92-100:81-88, 1993.

[2] I. Dagan, S. Marcus and S. Markovitch : Contextual word similarity and estimation from sparse data. *Computer Speech and Language*, 9:123-152, 1995.

[3] D. Hindle : Noun Classification From Predicate Argument Structures, Proc. of the 28th meeting of ACL, 1990.

[4] K. W. Church and P. Hanks : Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16:22-29, 1990.

[5] *Longman Dictionary of Contemporary English.* Longman, Harlow, Essex, new edition, 1987.