コーパスとシソーラスを利用した名詞間距離の設定

新納浩幸
茨城大学 工学部 システム工学科

1 はじめに

本論文では，一般の広い範囲の名詞の間の距離をコーパスと既存のシソーラスを用いて設定することを試みる．

自然言語処理の多くのアプリケーションで，名詞間の距離（あるいは類似度）を適切に求めることが必要とされている．特に，用例に基づく手法では，名詞間の距離が用例を選択する有益な尺度となっている [1, 2]．その他，並列句の解析 [3]，語彙的結束性の判定 [4] にも名詞間の距離が利用される．

名詞間の距離は，通常，シソーラスから計算される．ただし人手で作成するシソーラスは，機械処理に向いていない，構築コストが高いなどが指摘されており，コーパスを利用してシソーラスを自動構築する試みがなされている [5]．ただしこのようにして作成されるシソーラス中の名詞は，そのコーパスの分野に出現する名詞であり，一般の広い範囲の名詞に対するシソーラスは作成できない．これは，つき詰めれば，コーパス中に所望の名詞がほとんど出現しないというスパース性の問題と捉えることができる．

従来，スパース性に対処する方法としては，低頻度（あるいは未出現）の語の振舞いを，その語と類似の語の振舞いによって推定することが考えられてきた．

Brown は，ある単語列の次に現れる単語を予測するため，n-gram モデルを拡張した n-gram class モデルを提案した [6]．そこでは単語をクラスに変換することで，スパース性に対処している．Pereira も基本的には単語をクラスに変換することで，スパース性に対処している．ただし，単語はあるクラスに確率的に属するとしたソフトクラスターの概念を導入している [7]．Brown も Pereira も各々のモデルを基に，単語を妥当なクラスに割り当てるクラスタリングのアルゴリズムを与えている．Dangan は未出現の共起単語の相互情報量を，類似している語の共起単語の相互情報量から推定している [8]．更に同様のアイデアを基に，未出現の共起関係の確率も与えている [9]．

しかしこれらの研究では，類似の判断をコーパスから得られた情報により行なっている．そのためこの時点でスパース性の影響を受けていると思われる．また類似の語さえも未出現であるような場合には対処できない．

ここでは既存のシソーラスを利用して，未出現の語と類似している語を設定する．これによって，類似している語が存在しない場合でも，シソーラス上の類似性を利用して距離を設定できる．処理的には，まずコーパス中の名詞をシソーラス上のクラスに置き換え，クラス間距離を測る．次にコーパスだけでは設定できないクラス間距離をシソーラス上の類似のクラスの振舞いから推定する．類似している度合いを徐々に緩めることで，類似の語も未出現であるような語に対する距離が設定できてゆく．本手法で設定した距離は，利用したシソーラスと同規模のカバー率がある．しかもコーパスの分野性も反映していると考える．

本論文では，シソーラスとして分類語彙表 [10]，コーパスとして日経新聞 5 年分（約 785 万文）を利用して，分類語彙表に記載されている名詞（約 2 万 7 千種類）間の距離を設定した．

2 名詞間距離の設定

人手でシソーラスを作成する場合を考えてみると，非常に類似している名詞対は類似していると容易に判断できるが，類似度が低くなってゆくほど，類似しているかどうかの判断が困難になると思われる．つまりシソーラスは上位になるほど，クラス間の統合に信頼性がなくなり，逆にシソーラスの葉に近い部分は局所的に見ると信頼性が高いと思われる．

そこで，分類語彙表の葉に当たるクラス（ここではこれらを**原始クラス**と呼ぶ）内にある名詞は互いに類似している（距離が 0）という仮定をおく．すると，我々が求めるべきものは，原始クラス間の距離となる．原始クラス間の距離をコーパスから求める場合，コーパスのスパース性のために，あるクラス間の距離が定義できないという事態が生じる．この部分を原始クラス間の分類語彙表上の距離を利用して推定する．

本手法は基本的に以下の 4 つの step からなる．

step1 コーパスから共起データを収集する．
step2 共起データ中の名詞を原始クラスに置き換える．
step3 step2 から得られた共起データを利用して，原始クラス間の距離を測る．
step4 未定義となっている原始クラス間の距離を推定する．

2.1 共起データの収集 (step1)

本手法を適用するために，まず，コーパスから共起データを収集する必要がある．

コーパス中で名詞 A が格助詞 B を介して動詞 C と共起した場合に，[A，B，C] の 3 組を取り出す．この 3 組を共起データと呼ぶ．例えば「本を読んでいる」からは共起データとして [本，を，読む] が取り出せる．

コーパスから正確に共起データを取り出すことは一般に困難である．ここでは名詞，格助詞，動詞が以下のように連続して現れた場合のみを対象とする．

名詞 (N) + 格助詞 (R) + 動詞 (V)

このデータから [N，R，V] を取り出す．また本論文では，格助詞は「を」だけを対象にする．これは「を」格が名詞の分類を行なうのに最も適している格であるという経験的予測があるためである．

コーパスは日本経済新聞'90 年から'94 年の 5 年分（約 785 万文，1 文の平均文字数は 49.0 文字）を利用する．取り出した共起データの総数は約 441 万組 である．この中から，頻度が 1 であるもの，動詞の頻度が 20 未満であるものを取り除き，最終的に，総数 3,268,602 ，種類数 245,951 の共起データを作成した．このときの動詞は 3,577 種類であった．この共起データを以下の処理の対象にする．

2.2 クラスへの置き換え (step2)

step2 では，step1 で得られた共起データの名詞の部分を，その名詞が属する原始クラスに置き換える．

分類語彙表 [10] は最大 6 レベルの木構造になっており，木構造の葉の部分に，分類番号が割り振られている．各名詞は語義ごとに対応する分類番号を持っている．この分類番号が原始クラスに対応する．分類語彙表の原始クラスの総数は 3,582 である．

分類語彙表には複合名詞など，記載されていない名詞も多いので，共起データ中のすべての名詞が原始クラスに置き換えられるわけではない．また置換えの際には，多義語の問題が生じる．通常，名詞は多義であるために複数の原始クラスを持っている．ここでは同じ共起関係にある他の原始クラスの分布から，一部の多義は解消している．この処理の説明は割愛する．分類語彙表に記載されていない名詞を持つ共起データは step1 で得られた共起データ中，14.0% であった．また多義であったものは，17.5% であった．この多義のうち，17.6% の多義が解消された．最終的に得た共起データは，総数 2,708,135 ，種類数 115,330 であった．

2.3 コーパスからの測定 (step3)

step3 では原始クラス間の距離を設定する．ここでは Hindle が与えた名詞間の類似度を計算する手法を利用する [5]．

まず，動詞 vと原始クラス Cの相互情報量 MIを以下のように定義する．

$$MI(v,C) = \log_2 \frac{\frac{f(v,C)}{N}}{\frac{f(v)}{N}\frac{f(C)}{N}}$$

ここで，N は共起データの総数 (2,708,135)，$f(v), f(C)$ は共起データ中の v および C の総数，$f(v,C)$ は，共起データ [C, を,v] の頻度である．これは動詞 vに対するクラス Cの分布を表している．次に，動詞 vから見た，クラス C_i とクラス C_j の類似度 sim を以下のように定義する．

$sim(v, C_i, C_j)$

$$= \begin{cases} min(|MI(v,C_i)|, |MI(v,C_j)|) \\ \quad : MI(v,C_i) * MI(v,C_j) > 0 \\ 0 \quad : otherwise \end{cases}$$

最終的に，動詞全体から見た C_i と C_j の類似度を以下のように定義する．

$$SIM(C_i, C_j) = \sum_v sim(v, C_i, C_j)$$

上記の式で，v は step2 から得られた共起データ内の動詞 (3,577 種類) だけが対象なので，$f(v) > 0$ である．しかし原始クラス C は可能なすべての場合を考えるために，$f(C) = 0$ があり得る．$f(C) = 0$ の場合，$MI(v,C)$ が定義できない．そこですべての v に対して，$f(C_a) = 0$ あるいは $f(C_b) = 0$ の場合，$SIM(C_a, C_b)$ は未定義とする．

また距離は最終的に得られた類似度から以下のように定義する．

$$dis(C_i, C_j) = \frac{1}{SIM(C_i, C_j)}$$

2.4 シソーラスを用いた推定 (step4)

今，原始クラスは 3,582 種類あるので，${}_{3582}C_2 = 6,413,571$ 種類の距離を定義する必要がある．step3 によって，2,049,566 対の距離を定義することが出来た．これは全体の 32.0 % である．step4 では残り 68.0 % の未定義の距離を既存のシソーラスから推定する．

未定義となっているクラス C_a と クラス C_b 間の距離 $dis(C_a, C_b)$ を推定することを考える．まず C_a と直接の親ノードが共通であるような兄弟関係にあるクラスの集合 $\{C_{a_1}, C_{a_2}, \cdots, C_{a_i}\}$ を取り出す．C_bに対しても同様に$\{C_{b_1}, C_{b_2}, \cdots, C_{b_j}\}$ を取り出す．シソーラス上で類似度が大きい場合，その類似性は信頼性が高いので，定義済みの $SIM(C_{a_k}, C_b)$ あるいは $SIM(C_a, C_{b_m})$ は $SIM(C_a, C_b)$ に近いと考えられる．そこで定義済

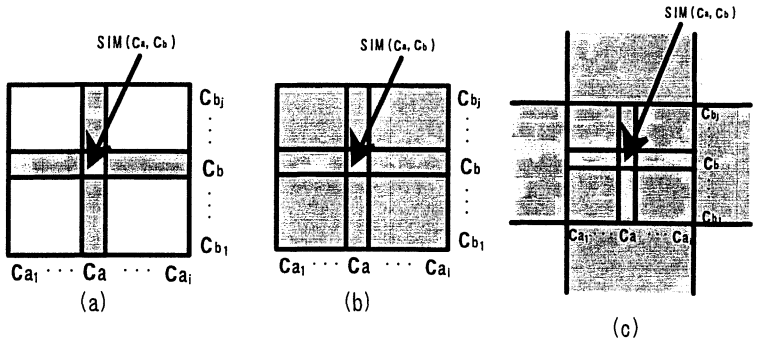

図 1: 未定義の距離の推定

みの $SIM(C_{a_k}, C_b)$ と $SIM(C_a, C_{b_m})$ の平均によって $SIM(C_a, C_b)$ を定義し，それにより $dis(C_a, C_b)$ を設定する．これは図 1 (a) の斜線の部分の平均で，目的の距離に対応する類似度を埋めることに相当する．

そしてこの操作を経ても，さらに未定義であるクラス間の距離がある場合，$SIM(C_{a_k}, C_{b_m})$ の平均によって定義する．これは図 1 (b) の斜線の部分の平均で，目的の距離に対応する類似度を埋めることに相当する．

さらに未定義であるクラス間の距離が残った場合は，祖父母のノードが共通であるようなクラスの集合をとりだし，先の操作を繰り返す（図 1 (c) 参照）．

ここでは図 1 (a) に対応する推定により，未定義部分の 75.3 % を推定でき，更に 図 1 (b) に対応する推定を行なうことで，未定義部分の 94.5 % を推定できた．更に 図 1 (c) に対応する推定を行なうことで，すべての未定義部分が推定できた．

3 評価

得られたクラス間の距離と分類語彙表上の距離とを比較してみる．分類語彙表上の距離は，対象のクラスの共通の親ノードの木構造上のレベルによって測ることにする．結果を表 1 に示す．

分類語彙表上の共通の親ノードの位置	得られた距離の平均
レベル 1	0.4630
レベル 2	0.2710
レベル 3	0.1534
レベル 4	0.0991
レベル 5	0.0675
レベル 6	0.0

表 1: 分類語彙表との距離の相関

分類語彙表上で距離が大きくなると，得られたクラス間の距離も大きくなる傾向がわかる．これにより，得られた距離は，大雑把に，分類語彙表上の距離の分布に沿っていることがわかる．

また第 1 段階の推定について，クラス間距離を設定するために使った距離の変動係数（標準偏差を平均で割った値）の平均は 0.384 であった．コーパスから得られた距離の変動係数が 2.125 であることから，推定値に用いた距離は互いに近い値になっていることがわかる．

次に動詞の語義選択の実験により，得られた名詞間の距離の妥当性を評価する．動詞の語義毎にいくつかの代表的名詞を用例として，テスト用の文の名詞とその用例との名詞との距離で，語義を選択する．この際，ある格の中の名詞の語義選択が，直接動詞の語義選択を決定するような動詞とその格（13 種類）を選んだ（表 2）．

距離としては本手法で得られた距離と，分類語彙表上の距離を用いる．ただし，分類語彙表上の距離は荒いために，候補が複数生じる場合がある．そのため分類語彙表上の距離の評価では，一意に正解を選べた場合には ○，候補の中に正解が含まれていた場合は △，候補の中に正解がなかった場合を × とした．また本手法での距離の評価では，もっとも距離が小さいものが正解の場合は ○，距離が 2 番目に小さいものが正解の場合は △，それ以外を × とした．結果を表 2 に示す．

本手法で得られた距離の方が，若干ではあるが，適切に語義を選択できていることがわかる．

4 考察

コーパスからの知識獲得において，すべての知識をコーパスから取り出すことは難しい．それは不完全な解析（特に多義性），コーパスのスパース性が原因である．このためどこかで人手の介入を効率良く行なうか，コー

動詞の パターン	語義 の数	用例の名詞	テスト用の問題	分類語彙表 ○	分類語彙表 △	分類語彙表 ×	本手法 ○	本手法 △	本手法 ×
を 起こす	9	27 (妹, 身体, 土, 会社, ...)	24 (看板, 母, 中毒, 水漏れ,...)	14	2	8	17	0	7
が 解ける	4	12 (ひも, 怒り, 暗号, 処分,...)	4 (結び, しこり, なぞ, 停学)	0	2	2	1	1	2
を 直す	7	25 (車, ネクタイ, 偏食, 誤字,...)	16 (道, ベンチ, 論文, 結核,...)	8	1	7	9	1	6
を 握る	5	14 (バット, 包丁, こぶし, 証拠,...)	3 (ひも, マイク, 成否)	1	1	1	1	0	2
に 乗る	5	18 (電車, 踏み台, 相談, リズム,...)	16 (ロケット, 椅子, 御輿, 人気,...)	14	0	2	13	2	1
に 触れる	3	16 (彼女, 問題, 核心, 愛,...)	8 (水, 地球, 社会, 制度,...)	7	0	1	7	0	1
を 味わう	3	13 (幸福, 酒, 古典, 名曲,...)	13 (感触, 余韻, ビール, 作品,...)	12	0	1	10	1	2
を 合わせる	5	13 (胸, 手, 話, 答え,...)	9 (力, 収入, 曲, つじつま,...)	3	1	5	3	1	5
を 押える	6	22 (帽子, 目, 怒り, 要点,...)	19 (弦, 犯人, アリバイ, 座席,...)	7	4	8	9	3	7
を 壊す	4	17 (家, 時計, 腹, 平和,...)	18 (戸棚, 肩, イメージ, 調和,...)	16	0	2	14	2	2
を 寝る	2	6 (計画, 対策, 刀, 飴,...)	8 (策, 構想, 考え, 粉,...)	7	0	1	8	0	0
を 許す	4	19 (帰宅, 彼, 夜勤, 心,...)	18 (建設, 参加, 利用, 浮気,...)	14	1	3	16	0	2
を 読む	4	19 (教科書, グラフ, 票, 顔色,...)	28 (解説, 日記, 声明, 動向,...)	13	6	9	16	3	9
Total			184	116	18	50	124	14	46

表 2: 語義選択の実験

パスとは別種の資源を利用することが必要になると思われる.

辞書の定義文からシソーラスを作成することや [11], 既存の知識から事例データを作成することも [12], 別種の資源の利用だと考えられる. そしてここでは別種の資源として既存のシソーラスを利用することを提案した. 人間の作った規則や知識は, 単純な事例よりもはるかに正確, 適切であり, ある意味で情報量が多い. これらを利用していくことが, スパース性への対処には有効だと考える.

また, 本手法は既存の知識と統計データを統合して, 利用しているとも考えられる. その場合, 本手法では統計データに重みをおいて, 類似度を設定しているが, 適切な重み付けの割合を調べることも必要である [12].

また既存のシソーラスでは分類が細か過ぎたり, 荒過ぎたりしている部分がある. シソーラスの木構造において, 原始クラスの1つ上位にあるレベルで, 原始クラスをまとめたクラスを考えてみると, そのクラス内の原始クラス間距離の平均が小さいものほど, そのクラスの分類が細かく, 大きいほど荒いことが推測できる. この点から既存のシソーラスの改良も可能だと思われる. 本来, ここでの手法は既存のシソーラスの修正という位置づけで捉えることもできる.

5 おわりに

本論文では, 名詞間の距離をコーパスから自動設定する際に生じるコーパスのスパース性の問題に対処するために, 既存のシソーラスを利用した.

ここで設定できた距離は既存のシソーラスと同規模のカバー率を持つ. またコーパスの分野性も反映できている.

コーパスのスパース性に対処するには, 既存の資源(知識)を活用することも有効だと考える. 今後はこの実験を通して得られた共起データと名詞間距離を基に, 既存シソーラスの修正, 拡張を行ないたい. また動詞の語義分類も試みたい.

謝辞

本実験で利用したコーパスは, 日本経済新聞 CD-ROM '90 ～ '94 版から得ています. コーパスの利用を許可して頂いた日本経済新聞社, 及び, このコーパスの利用に関して尽力された方々に深く感謝します.

参考文献

[1] 佐藤理史 : "MTB1: 実例に基づく訳語選択", 人工知能学会誌,Vol.6,No.4,pp.592–600(1991).

[2] 佐藤理史 : "MTB2: 実例に基づく翻訳における複数翻訳例の組合せ利用", 人工知能学会誌,Vol.6,No.6,pp.861–871(1991).

[3] Kurohashi,S., and Nagao,M. : "A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures", Computational Linguistices,Vol.20,No.4,pp.507–534(1994).

[4] Okumura,M.,and Honda,T. : Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of COLING-94*, pp.755–761(1994).

[5] Hindle,D. : Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pp.268–275(1990).

[6] Brown,P.F.,Pietra,V.D, deSouza,P.V., Lai,J.C. and Mercer,R.L. : "Class-Based n-gram Models of Natural Language", Computational Linguistics, Vol.18,No.4,pp.467–479(1992).

[7] Pereira,F., Tishby,N., and Lee,L. : "Distributional clustering of English word", In *Proceedings of ACL-93*, pp.183–190(1993).

[8] Dangan,I., Marcus,S., and Markovitch,S. : "Contextual word similarity and estimation from sparse data", In *Proceedings of ACL-93*, pp.164–171(1993).

[9] Dangan,I., Pereira,F., and Lee,L. : "Similarity-Based Estimation of Word Coocurrence Probabilities", In *Proceedings of ACL-94*, pp.272–278(1994).

[10] 国立国語研究所 : "分類語彙表", 秀英出版 (1994).

[11] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将 : "国語辞書を用いたシソーラスの作成について", 情報処理学会自然言語処理研究会, Vol.91,No.37,91-NL-83, pp.121–128(1991).

[12] 金田重郎, 秋葉泰弘, 石井恵 : "事例に基づく英語動詞選択ルールの修正型学習手法", 言語処理学会第1回年次大会予稿集,pp.333–336(1995).