検索エンジンを利用した未登録単語に関する単語間距離の測定

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1 はじめに

多くの自然言語処理システムでは，単語間の意味的な距離（あるいは類似度）を測る処理を，本質的に必要とする．例えば，事例ベースの手法では事例間の距離を測る処理が必須であり，その処理の核には単語間距離の測定がある．また，単語間距離を測ることができれば，シソーラスを自動構築できる．これは，シソーラスを利用したシステムでは，シソーラスを使う部分を単語間距離を測る処理に置き換えられることを意味している．つまり単語間距離の測定は自然言語処理の重要な要素技術と言える．

単語間距離を測るには，通常，用意してあるデータベース（例えばシソーラスなど）を用いて測定される．しかしデータベースは静的であり，必ず未登録単語が存在し，未登録単語に関する単語間の距離は測定できないという問題がある．つまり，単語 U をデータベースに登録されていない単語（未登録単語），単語 K を登録単語（あるいは未登録単語）としたとき，U と K の距離 $d(U,K)$ を測ることはできない．これはシソーラスを利用したシステムの多くで問題となる未知語の問題と本質的に同じである．

ここではこの問題に対処することを目的に，検索エンジンを利用して $d(U,K)$ を測ることを試みる．まず Web に適した基底の単語 $b_1, b_2, \cdots, b_n$ を選出し，“$U\ b_i$” というキーワードから検索エンジンを利用して共起したページ数 f_i を得る．$(f_1, f_2, \cdots, f_n)$ を正規化したものを U の R^n 上の座標とする．同様にして K の座標も得られるので，$d(U,K)$ の距離を測ることができる．

実験では，未知語になりそうな単語を選び，その単語に関して，同類の単語 2 個（A 群）と似ているが少し意味が異なる単語 2 個（B 群）を選出し，それら 4 単語と未知語間の距離を測った．未知語と最も類似した 2 つの単語が A 群の 2 つであることを正解とするという評価方法をとった．この実験を 30 個の未知語に対して行なった．30 個中 11 個の正解を得た．また適当に選出した 25 単語からコーパスを利用したクラスタリング，本手法を用いたクラスタリング，両者を併用したクラスタリングを行った．両者を併用したクラスタリングではかなりよい結果が得られた．これらの結果から，本手法は未登録語に限定して使うのが適当であると考える．基底の単語の選出方法を今後の課題とする．

2 検索エンジンを利用した単語間距離

2.1 コーパスを利用した単語ベクトルの作成

単語 U が R^n 上の座標（n 次元ベクトル）として表現できれば，単語間の距離を測ることができる．問題はどのようにして U を n 次元ベクトルで表現するかである．

通常，適当な単語列 $b_1, b_2, \cdots, b_n$ を選出し，U と b_j の共起頻度 f_j をコーパスから得て，$\{f_j\}$ を正規化することで U のベクトルを得る．本論文では単語列 $b_1, b_2, \cdots, b_n$ を 基底単語列 と呼んでいる．

ただしこの作成方法は，基底単語列をどのように作成するかという問題[1]の他に，コーパスのスパース性の問題がある．つまり，頻度の低い U に関しては，U と b_j の共起頻度がほぼ 0 であり，適切なベクトルが得られない．この問題は，U に関する単語間距離を測定できないという問題を生じさせる．

2.2 検索エンジンを利用した単語ベクトルの作成

上記の問題の解決として，Web をコーパス代わりにすることが考えられる．Web 上のテキストは巨大であり，未知語は存在しないと仮定できる．

ただし共起頻度を測ることが実質不可能である．このため，ここでは共起の定義を「同じページ内で現れる」に変更する．その場合，既存の検索エンジンを利用すれば，そのヒット件数から共起頻度が得られ，その結果，単語ベクトルが得られることになる．

[1] 次章で述べる．

3　基底単語列の作成

基底単語列は意味的には単語ベクトルが構成する空間を n 次元だと考え，その第 i 次元の単位ベクトルに相当する単語 e_i を並べたものである．意味素と呼ばれるものに相当する．

ただし，実際は何が「単位ベクトルに相当する単語」であるかは分からないし，そのような単語が存在するのかどうかもはっきりしない．LSI のように，数理的に既存の単語列から，単位ベクトルに相当するベクトル u_i を求めることも行われているが [6][2]，ここでは距離測定を容易にするために，単語をクラスタリングすることで基底単語列を作成した．つまり，クラスタリングによりクラスターが n 個できる．第 i クラスターから代表の単語 b_i 選び，b_i の列を作ることで，基底単語列を作成する．このように作成した基底単語列を作成した場合，b_i と b_j は全く異なる意味の単語であることから，直交している，$\{b_i\}$ は空間を張っている，というイメージで捉えることができ，基底単語列をなす条件をほぼ満たしていると考えられる．

実際に作成した手順は以下の通りである．

論文 [2] ではクラスタリングにより名詞単語を 922 クラスターに分類している．この各クラスターから適当に 1 単語取り出し，922 個の単語の列 $\{a_i\}$ を作る．これを直接，基底単語列とすることも考えられるが，このクラスタリングは新聞記事コーパスを利用して行っているので，この $\{a_i\}$ では Web の単語の空間を張っているとは言えない．

そのために $\{a_i\}$ を Web のデータを用いてクラスタリングした．具体的には，各 a_i をキーワードとして Google で検索し，ヒットした件数 f_i を得る．次に a_i と a_j を組にした “a_i a_j” というキーワードにより，Google でのヒット件数 f_{ij} を得る．そして a_i と a_j の類似度を以下のダイス係数で定義する．

$$sim(a_i, a_j) = \frac{2 \cdot f_{ij}}{f_i + f_j}$$

各単語間に類似度が定義できたので，クラスタリングが行える．ここでは群平均法 [1] を利用した．

いくつのクラスターに分けるのがよいかは難しい問題である．ここでは適当に 100 クラスターにした．最後に各クラスターから Web データ中の頻度が最も低い単語を取り出すことで基底単語列を作成した．実際の単語列を以下に示す．

以降，ＰＣ，完成，対策，音声，成功，スタート，流れ，コーヒー，展開，話し，保護，周囲，資格，向こう，ケース，空気，マーク，メーリングリスト，半分，電源，翌日，半年，アドバイス，お客さん，実家，建設，公園，講師，世界中，ヒット，誕生日，職場，エン，安定，ガン，性能，オフ，以後，影響，クリスマス，案内，願い，伝統，教員，個性，エンジン，解説，アルバム，苦労，以来，ゴミ，価値，中心，デジカメ，ルール，メイン，コンピューター，スト，要求，外国人，野球，被害，残り，違い，本物，事情，理論，職業，スペース，高知，事務所，教室，回答，世代，人類，用意，メモ，固定，範囲，通り，社長，宿泊，感覚，先輩，ワイン，中身，効率，新宿，田舎，この後，事態，都道府県，記録，男女，前述，青森，冗談，ご存知，積極

4　実験

4.1　未登録単語に関する距離測定

以下のキーワードランキングサイトの上位 30 個の単語[3]を未登録単語と考えた．

http://guide.search.goo.ne.jp/ranking/

各キーワード w に対して，それと意味的に近いキーワード α_1 と α_2，似ているが意味が異なるキーワード β_1 と β_2 を作成し，合計 5 個のキーワードのセット $\{w, \alpha_1, \alpha_2, \beta_1, \beta_2\}$ を作る．つまり，このセットが 30 組作成される．各組に対して，提案手法により w と他の単語との距離を測定し，w から近い順に $\{\alpha_1, \alpha_2, \beta_1, \beta_2\}$ を並べる．1 番目に近い単語と 2 番目に近い単語が，α_1 か α_2 であった場合に，その組の検査は正解と判定する．

実験の結果を表 1 に示す．1 列目が未知語，2，3 列目が類似語，4，5 列目が非類似語，6 列目の括弧が実験結果である．また各単語の後の括弧（[ ]）内の数値は対象単語と近い順の順位を示す．

結果 30 組中，11 組が正解であった．ランダムに並べた場合に正解する確率は，$(2+2)/4! = 1/6$ なので，30 組をテストすると平均 5 個しか正解が得られない．このため，本手法は「何もしないよりも効果はある」ことが分かる．

[2] 基底単語ではなく，基底のベクトルである．

[3] 2005 年 11 月 29 日前後のものである．

表 1: 未登録単語に関する距離測定

未登録単語	類似単語1	類似単語2	非類似単語1	非類似単語2	判定
三井住友銀行	UFJ[2]	みずほ銀行 [1]	野村證券 [3]	大和證券 [4]	○
十支	犬 [2]	鳥 [1]	猫 [3]	象 [4]	○
バレーボール	バスケットボール [1]	サッカー [4]	柔道 [3]	剣道 [2]	×
JR 東日本	JR 西日本 [1]	JR 東海 [2]	JAL[3]	ANA[4]	○
朝日新聞	読売新聞 [1]	毎日新聞 [4]	日刊スポーツ [2]	スポーツ報知 [3]	×
NHK	TBS[2]	フジテレビ [3]	J-WAVE[1]	bayfm[4]	×
オンラインゲーム	ハンゲーム [2]	ファイナルファンタジー [1]	ベー駒 [3]	けん球 [4]	○
集英社	講談社 [2]	小学館 [1]	凸版印刷 [3]	大日本印刷 [4]	○
セブンイレブン	ローソン [1]	サンクス [3]	西友 [4]	イオン [2]	×
ダイエー	イトーヨーカ堂 [2]	パルコ [1]	セシール [3]	ディノス [4]	○
年賀状	書中見舞い [2]	招待状 [3]	宅急便 [1]	小包 [4]	×
浜崎あゆみ	中島美嘉 [2]	大塚愛 [1]	原田泰造 [4]	太田光 [3]	○
チャングムの誓い	美しき日々[1]	冬のソナタ [2]	あずみ [3]	座頭市 [4]	○
バイオハザード 4	ドラゴンクエスト [2]	スーパーマリオ [1]	ビートマニア [4]	ムシキング [3]	○
ドラゴンボール	ワンピース [1]	NARUTO[4]	ライオンキング [3]	美女と野獣 [2]	×
シャープ	ソニー [2]	パナソニック [4]	ヤマダ電器 [1]	ヨドバシカメラ [3]	×
あいのうた	花より男子 [3]	木更津キャッツアイ [2]	ハリーポッターと炎のゴブレット [1]	エンパイア・オブ・ザ・ウルフ [4]	×
B'Z	WaT[2]	ゆず [3]	ロバート [1]	おぎやはぎ [4]	×
仮面ライダー響鬼	シャイダー [3]	ギャバン [4]	ケロロ軍曹 [2]	トランスフォーマー [1]	×
電車男	祖母の肖像 [4]	バカみたい [2]	夏のレプリカ [3]	ジョーカー [1]	×
スタッドレスタイヤ	ホイール [2]	ワイパー [1]	サーフボード [3]	ウェットスーツ [4]	○
リュ・シュウォン	イ・ビョンホン [2]	ペ・ヨンジュン [3]	オダギリジョー [1]	細川茂樹 [4]	×
F1	ラリー [1]	フォーミュラニッポン [3]	モトクロス [4]	オートレース [2]	×
ハローワーク	人材銀行 [2]	労働基準監督署 [3]	九州厚生局 [4]	東北厚生局 [1]	×
ぐるなび	グルメウォーカー [1]	グルコン [3]	楽天デリバリー [4]	すぐくるデリバリー [2]	×
タウンページ	電話帳 [1]	ハローページ [2]	郵便番号 [3]	住所コード [4]	○
気象庁	国土地理院 [2]	海上保安庁 [3]	会計検査院 [1]	人事院 [4]	×
ホテル	民宿 [3]	旅館 [4]	クルーズ [2]	ハネムーン [1]	×
ロト 6	ナンバーズ [1]	ミニロト [4]	toto[2]	パチスロ [3]	×
ぷらら	OCN[4]	So-net[1]	Google[2]	JWord[3]	×

4.2 コーパスとの相補的利用

我々は以前，適当に選出した単語 25 個をコーパス（毎日新聞 '95 年度記事 1 年分）を利用してクラスタリングを行った [4]．以下がその 25 単語である．

動物 (a1) プードル，(a2) チワワ，(a3) 犬，(a4) 猿，(a5) ゴリラ

食べ物 (b1) カレー，(b2) ラーメン，(b3) スパゲッティー，(b4) 焼きそば，(b5) ハンバーグ

感情や人生観 (c1) 幸福，(c2) 満足，(c3) 愛情，(c4) 結婚，(c5) 運命

繁栄しているもの (d1) 情報，(d2) 知識，(d3) 手段，(d4) 交通，(d5) 設備

地名 (e1) 今帰仁，(e2) 沖縄，(e3) ブリスベン，(e4) オーストラリア，(e5) オーストリア

また図 1 がそのクラスタリング結果である．括弧内の数値はコーパス中のその単語の頻度を示す．この結果からコーパスのスパース性が悪影響を及ぼしていることが分かる．

本手法を用いて，上記 25 単語を再び，クラスタリングすると，図 2 の結果が得られた．

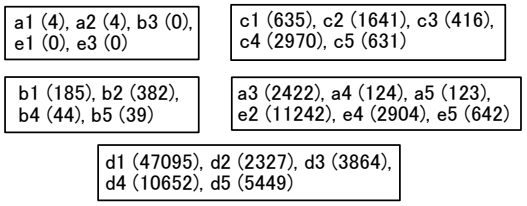

図 1: コーパスを使ったクラスタリング

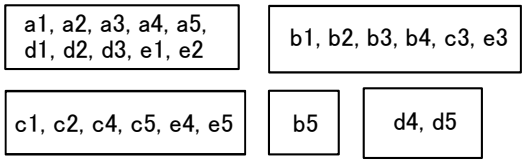

図 2: 検索エンジンを使ったクラスタリング

コーパスを使ったクラスタリングほどうまくクラスタリングできているようには見えない．ただし，コーパス中で頻度が低かった単語,「スパゲッティー」「今帰仁」「ブリスベン」「プードル」「チワワ」に対して，その他の単語で最も距離が近い単語を，本手法により

求めると以下の結果を得た．

```
(b3) スパゲッティー    →  (b2) ラーメン
(e1) 今帰仁            →  (e2) 沖縄
(e3) ブリスベン        →  (b2) ラーメン   (*)
(a1) プードル          →  (a3) 犬
(a2) チワワ            →  (a1) プードル
                          次は (a3) 犬
```

類似性の判定は (*) の１つを除いて適切である．

上記の実験結果から，コーパスのスパース性の影響がある部分だけに対して，本手法を用いることが考えられる．具体的には，スパース性のある単語を省いてコーパスによりクラスタリングし，次に本手法を用いて，省かれた単語をクラスターに割り当ててゆく方法である．この方法で得られたクラスタリング結果を以下に示す．丸で囲まれたものは，本手法によりクラスターに割り当てられた単語である．ほぼ正しいクラスタリングが得られている．

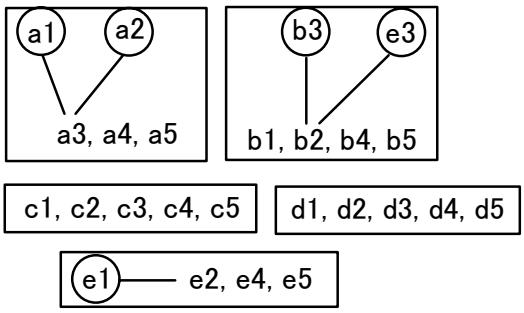

図 3: 本手法を補助的に利用したクラスタリング

5　考察

本手法が有効に機能するかどうかは，基底単語列の選出にかかっている．基底単語列を別のものに設定すると，全く異なる結果が得られる．ここでは 922 個の単語を 100 個のクラスターにクラスタリングしたが，他の個数（922 個，50 個，25 個）も試してみた．しかし，ここで報告した以上の結果は得られていない．更にクラスターから１つの単語を選出する部分でも，その選出の方法を変更しただけで結果が異なってくる．ここでは Web データ中で頻度の最小のものを選出した．最大のものを選出する方法も試したが，これも，本報告以上の結果は得られていない．結論的には，そこそこうまくいくような基底単語列がたまたま見つかったとも言えるかもしれない．基底単語列の選出方法については更に調査研究が必要である [5]．

本手法だけで単語間距離を測るには，その計り方はおおざっぱすぎる．そのため未登録語ではない一般の単語に対してまで本手法を用いるべきではない．前章の実験でも，そのことを裏付けている．コーパスにも現れないような未登録単語に関する距離を測る場合，既存の手法ではまともな対処方法がない．そのような状況では，本手法は有益だと考える．

また副次的な効果であるが，本手法は単語だけではなく，複合語などの表現に対する距離を測ることができる．Google はキーワードとして様々な表現を受け付けるからである．例えば，ある作品のタイトルを与えた場合に，それが映画のタイトルなのか，本のタイトルなのか，テレビ番組のタイトルなのかなどが判定できる可能性がある．現在 Web ページの分類に利用することを検討している [3]．

6　おわりに

本論文では，未登録単語に関する距離を測れないという従来の問題に対処するために，Web の検索エンジンを利用して単語間の距離を測ることを提案した．手法としては，基底の単語との AND 検索により得られるヒット件数から単語のベクトル表現を得ている．実験では，ある程度の精度を示した．本手法は既存の知識の補助として利用するのが適切だと考える．基底単語列の作成方法が今後の課題である．

参考文献

[1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, Vol. 31, No. 3, 1999.

[2] Minoru Sasaki and Hiroyuki Shinnou. Automatic thesaurus construction using word clustering. In *PACLING-2003*, pp. 55–62, 2003.

[3] 佐々木稔, 新納浩幸. 文書分類手法を用いた企業 Web サイトからの業種分類. 言語処理学会第 12 回年次大会, p. to appear, 2006.

[4] 大城亜里砂, 新納浩幸, 佐々木稔. 検索エンジンを利用した単語クラスタリング. 言語処理学会第 10 回年次大会, pp. 17–20, 2004.

[5] 藤井丈明, 新納浩幸, 佐々木稔. Web における基底単語の選出. 言語処理学会第 11 回年次大会, pp. 45–48, 2005.

[6] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2001.