Web ディレクトリを利用した名詞のジャンルベクトルの作成

林華 ○
茨城大学工学部情報工学科

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

本論文では名詞間の類似度を測ることを目的に、名詞に対するジャンルベクトルを構築する手法を提案する。

自然言語処理にとってシソーラスが有用であることは明らかである。用例ベースの翻訳などに見られるように、シソーラスの主な用途は名詞間の類似度の測定である。しかしその測定方法はシソーラスを表す木構造の位置関係から類似度を測るという単純なものであるため、荒い類似度しか得ることができない。本論文では名詞をジャンルベクトルと呼ばれるベクトルで表現する。これによって各名詞間に細かい類似度を与えることができる。

名詞に通常の形態情報の他にジャンルの情報を与えることで、より深い意味解析を行うことができるが [2]、本論文のアイデアは名詞に１つのジャンルを与えるのではなく、設定した各ジャンルとその名詞との関連度を与えるというものである。つまり、ジャンルベクトルとはベクトルの各次元をジャンルに設定し、名詞とそのジャンルとの関連度をその次元の値としたものである。名詞をジャンルベクトルで表現する場合、どのようなジャンルを設定するか、名詞とそのジャンルとの関連度をどのように求めるかという２つの問題がある。本論文では Web ディレクトリを利用して上記２つの問題を解決する。

実験では、名詞クラスタリングと文書クラスタリングに構築した名詞のジャンルベクトルを利用し、その有効性を示した。

2　名詞のジャンルベクトル

ジャンルを K 個設定する。i 番目のジャンルを j_i とする。名詞 a のジャンルベクトルとは K 次元のベクトルであり、第 i 次元の値 a_i は名詞 a とジャンル j_i との関連度である。

$$a = (a_1, a_2, \cdots, a_K)$$

またベクトルは正規化されている、$|a| = 1$。

名詞 a と名詞 b の類似度 $sim(a, b)$ は余弦で定義する。a と b は正規化されているので、類似度は a と b の内積 $a \cdot b$ となる。

$$sim(a, b) = \frac{a \cdot b}{|a||b|} = a \cdot b$$

名詞のジャンルベクトルを作成する際の問題は、j_i の設定と a_i の算出である。本論文では Web ディレクトリを利用して、これらの問題を解決する。

2.1　Web ディレクトリを利用したジャンルの設定

Web ディレクトリとは Web 上のホームページを階層的に分類したものである。基本的には木構造になっており、各ノードが分類のタイトル、リーフノードが各ホームページになっている。

本論文は Yahoo! の Web ディレクトリを利用する。図１がその Web ディレクトリのトップページである。

図１はルートノードの下の第１階層のノードのタイトルが並び、その下に第２階層のノードのタイトルが複数個示されている。

ここで最初のノードの「エンターテインメント」をクリックすると、「エンターテインメント」の１つ下位のノードが表示される（図２）。これらのノードが第２階層である。

以下同様に各階層のノードが表示され、最終的には指定された分類に属するホームページのタイトルが表示される。このタイトルはそのホームページにリンクされている。

本論文では第２階層に属するノードを各ジャンルに設定する。ジャンルの種類数は 362 である。このため本論文のジャンルベクトルの次元は 362 となる。

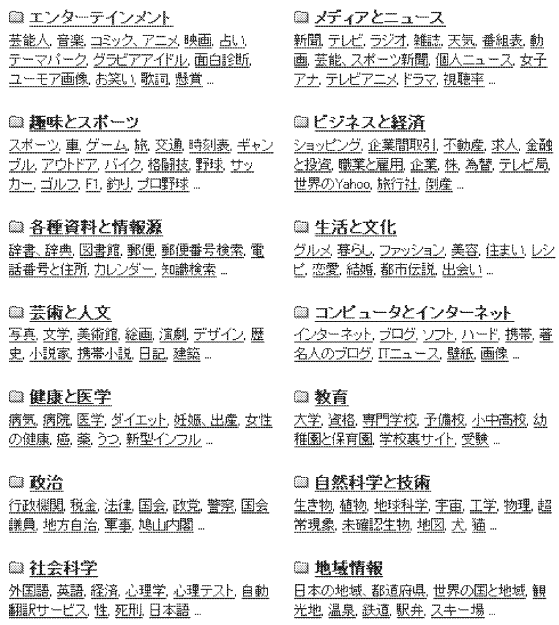

図 1: Web ディレクトリ

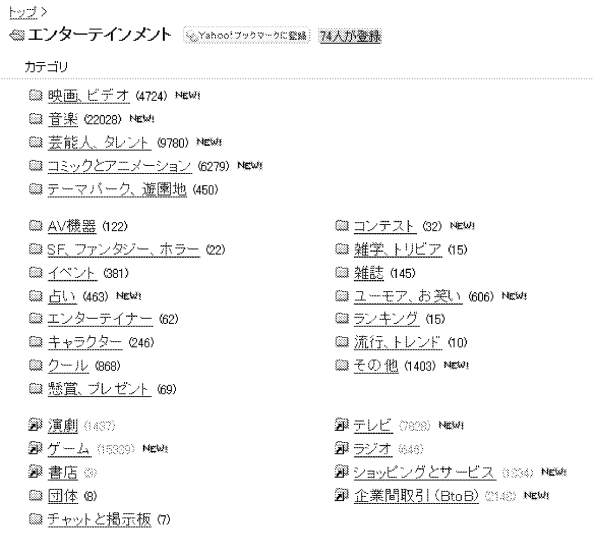

図 2: 「エンターテインメント」の第 2 階層

2.2 ジャンルの種名詞

名詞 a のジャンルベクトルの第 i 次元の値 a_i は名詞 a と第 i 番目のジャンル j_i との関連度である。a_i の値を求めるために、ジャンル j_i と関連度の高い名詞をいくつか用意しておき、それをジャンルの種名詞と呼ぶことにする。次節で述べるが、a_i の値は基本的に、ジャンル j_i の種名詞との関連度から計算する。

本節では種名詞の設定法を述べる。ここでも Web ディレクトリを利用する。本論文では Yahoo! の Web ディレクトリの第 2 階層に属するノードを各ジャンルに設定している。基本的にはジャンルのノードのサブジャンルの名称を利用する。Yahoo! の Web ディレクトリの第 1 階層の「エンターテイメント」のサブジャンルである「映画、ビデオ」の下には以下のようなサブジャンルが並んでいる。これらは Web ディレクトリの第 3 階層に属するノードとなっている。

> 「地域情報」「ジャンル」「日本映画」「外国映画」「作品」「映画館」「映画制作」「脚本」「監督」「俳優、女優」「映画祭」「特集上映」「試写会」「予告編」「映画音楽」「映画史」「テレビ番組」「ホームシアター」「イベント」「グッズ」「クラブ」「雑誌」「賞、コンテスト」「大学映画研究会」「団体」「チャットと掲示板」「データベース」「評論、レビュー」「リンク集」「配給会社」「書店」「ビデオカメラ、ビデオデッキ」「映画関連企業」「ビデオ関連企業」

基本的に上記の単語列から名詞をとりだしたものがジャンル「エンターテイメント→映画、ビデオ」の種名詞となる。

ジャンルの名称には種名詞として明らかに適切でないもの（例えば「リンク集」）や、単純な名詞でないもの（例えば「ビデオ関連企業」）があるので、そのようなものは機械的あるいは若干の手作業で取り除いたり整形したりした。

2.3 名詞とジャンルとの関連度計算

ジャンル j_k の種名詞を $\{n_1^{(k)}, n_2^{(k)}, \cdots, n_h^{(k)}\}$ とする。名詞 a と名詞 b の関連度を $r(a,b)$ で表すことにし、$r(a, n_j^{(k)})$ $(j = 1, 2, \cdots, h)$ の中から値の大きなもの 15 個をとり、それを $r(a, n_{j1}^{(k)})$,$r(a, n_{j2}^{(k)})$, $\cdots$,$r(a, n_{j15}^{(k)})$ とする。そして a_i はそれらの和で定義する。

$$a_i = r(a, n_{j1}^{(k)}) + r(a, n_{j2}^{(k)}) + \cdots + r(a, n_{j15}^{(k)})$$

名詞 a と名詞 b の関連度 $r(a,b)$ はコーパス[1]から求める。名詞 a のコーパス中の頻度を $f(a)$ で表す。名

[1] ここではコーパスとして BCCWJ コーパス [1] の「白書」を利用した。

詞 a と名詞 b が同一文内に共起した回数を $f(a,b)$ で表す。$r(a,b)$ は以下の dice 係数で定義する。

$$r(a,b) = \frac{2f(a,b)}{f(a)+f(b)}$$

以上より a_i が計算でき、名詞 a のジャンルベクトルが作成できる。

3 実験

作成した名詞のジャンルベクトルを評価するために、ここでは名詞クラスタリングと文書分類の 2 つの実験を行った。

3.1 名詞クラスタリング

名詞クラスタリングでは以下の 25 個の名詞をクラスタリングする。

「日本」「自民党」「野球」「リンゴ」「親切」「茨城」「予算」「テニス」「ご飯」「怒り」「日立」「献金」「バレー」「にんじん」「愛」「広島」「議員」「大相撲」「野菜」「思いやり」「アメリカ」「民意」「サッカー」「漬物」「努力」

これら名詞を 5 つのクラスタに分類する。正解は以下とした。

クラスタ 1 「日本」「茨城」「日立」「広島」「アメリカ」
クラスタ 2 「自民党」「予算」「献金」「議員」「民意」
クラスタ 3 「野球」「テニス」「バレー」「大相撲」「サッカー」
クラスタ 4 「リンゴ」「ご飯」「にんじん」「野菜」「漬物」
クラスタ 5 「親切」「怒り」「愛」「思いやり」「努力」

それぞれの名詞を本論文で示したジャンルベクトルで表現し、クラスタリングを行った。クラスタリングにはクラスタリングツールの CLUTO を利用した。

ここで CLUTO について注記しておく。CLUTO は強力なクラスタリングツールであり、`http://glaros.dtc.umn.edu/gkhome/views/cluto` で公開されている。CLUTO はクラスタリング手法や類似度関数を様々に設定できるが、ここでは default である k-way clustering と呼ばれる手法を用いた。また類似度としては直接類似度行列を与えて scluster で実行した。

得られた結果は以下である。エントロピーは 0.256、純度は 0.760 であり、良い結果が得られている。このことから本論文で示したジャンルベクトルがある程度妥当であることがわかる。

クラスタ 1 「大相撲」
クラスタ 2 「自民党」「献金」 「議員」「民意」
クラスタ 3 「野球」「テニス」「バレー」「サッカー」
クラスタ 4 「リンゴ」「ご飯」「にんじん」「野菜」「漬物」
クラスタ 5 「日本」「茨城」「日立」「広島」「アメリカ」「親切」「怒り」「愛」「思いやり」「努力」「予算」

3.2 文書クラスタリング

利用する文書データは Web のニュース記事 394 文書である。これは 2003 年 11 月 25 日から 12 月 5 日までの 10 日間でニュースサイト `http://news.goo.ne.jp/` に掲載されたニュース記事である。5 カテゴリ (政治、経済、国際、社会、スポーツ) から集めた。この文書データに対してクラスタリングを行う。

文書はベクトル空間モデルからベクトル化され、そのベクトルを利用して文書間の類似度が計算される。この類似度を $s1$ としておく。次に文書 d 中に名詞 a が h_a 個存在したとする。a がジャンルベクトルであることに注意して、文書 d を以下のベクトルで表現する。

$$d = \sum_{a \in d} h_a a$$

これを正規化したものを文書 d のジャンルベクトルと考える。このジャンルベクトル間の類似度からも、文書間の類似度が計算される。この類似度を $s2$ とする。最終的に文書間の類似度は $s1 + s2$ で定義する。

この類似度に従って、上記の文書データに対して CLUTO を用いてクラスタリングを行った。エントロピーは 0.601、純度は 0.539 であった。

比較のために $s1$ の類似度だけを用いた場合、エントロピーは 0.662、純度は 0.514 であり、ジャンルベクトルを利用する効果が確認できる。

またここでは $s2$ を名詞のジャンルベクトルから設定したが、シソーラスを利用して $s2$ を設定することもできる。ここではシソーラスとして分類語彙表を用いる。まず分類語彙表から得られる名詞の分類番号をすべて列挙する。K 種類あるとすると、名詞 a を K 次元のベクトルで表現できる。一般に名詞 a は複数の語義を持つので、それぞれの語義に対応する分類番号の次元の値を 1 に、残りは 0 とした K 次元のベクトルである。これを正規化して $|a| = 1$ とする。名詞 b が K 次元のベクトルであれば、ジャンルベクトルのときと同様に、文書 d の名詞 b の頻度が h_b であるとき、文書 d を以下のベクトルで表現できる。

$$d = \sum_{b \in d} h_b b$$

これを正規化したものを文書 d のシソーラスベクトルと考える。このシソーラスベクトル間の類似度から文書間の類似度が計算される。これを $s2$ に設定する。最終的に文書間の類似度は $s1 + s2$ で定義して、上記と同じ文書データでクラスタリングを行った場合、エントロピーは 0.654、純度は 0.557 であった。

この実験結果から、シソーラスを利用する効果はあったが、ジャンルベクトルの方がシソーラスよりも的確な類似度を与えていることがわかる。

4 考察

文書クラスタリングの実験では文書間類似度は $s1 + s2$ となっており、$s1$（通常の文書ベクトルからの類似度）と $s2$（文書のジャンルベクトルからの類似度）とを同じ重みで取り扱っている。$s1$ の重みを 1、$s2$ の重みを 0 にした場合、これは通常の類似度になるが、この場合、$s1 + s2$ の方が良い結果となっている。また $s1$ の重みを 0、$s2$ の重みを 1 にした場合、文書のジャンルベクトルからの類似度のみを使うことを意味するが、この場合は、上記文書クラスタリングの実験で、エントロピーは 0.707、純度は 0.506 となり、$s1$ の場合よりも悪い結果となっている。

これは文書のジャンルベクトルが十分に適切になっていない、つまり構築した名詞のジャンルベクトルが十分に適切になっていないことを示唆している。文書のジャンルベクトルは通常の文書ベクトルを次元縮約したものと見なせる。そのため 2 つの文書に同じ単語が現れない場合、通常、その文書間の類似度は 0 となるが次元縮約したベクトルを使うと、類似度が出てくる。その類似度が適切となり、$s1$ よりも良い結果になることを期待したが、そのような結果にはならなかった。ただし $s1 + s2$ が $s1$ よりもよいことは、ある程度は名詞のジャンルベクトルが適切であることを示唆している。

文書のジャンルベクトルの各次元があるジャンルとの関連度を表していると見なせるので、ジャンルベクトル自体からクラスタリングの結果を得ることができる。本論文ではジャンルベクトルの次元は 362 であり、これは Web ディレクトリの第 2 階層にあたり、第 1 階層まで次元を縮約すると次元は 14 となる。つまり 14 のクラスタに自動的に分割できる。実験では文書データを 5 つのクラスタに分割しているので、直接の比較はできないが、ジャンルベクトル自体を観察すると、この方法でクラスタリングしたり文書分類を行ったりすることはできないと思われる。これは名詞のジャンルベクトルから名詞をクラスタリングしたり名詞の語義識別したりするのが、適切に行えないことも意味している。この原因は名詞のジャンルベクトルのある特定の次元だけが常に高い値を取る傾向が見られるからである。これは名詞のジャンルベクトルの作成法に原因がある。あるジャンルの種名詞が多くの単語と関連度が高くなるケースがあることから生じている。

現在、作成できている名詞のジャンルベクトルを改良するには、種名詞の設定が鍵だと思われる。本論文では種名詞を自動で得るために Web ディレクトリを利用したが、種名詞の個数は少なくてもよいと思われる。実際に本論文でも名詞とジャンルとの関連度を測る際に、対象名詞と関連度の高い上位の 15 個の種名詞しか利用していない。当初、全ての種名詞を利用した実験も行ったが、種名詞の数がジャンルによりかなりばらつきがあり、対象単語と種名詞との関連度の平均をとった場合、ジャンルとの関連度が適切に得られてはいなかった。種名詞の個数は少なくてもよいと思われるので、今後は手作業で種名詞を選定して、名詞のジャンルベクトルを改良したい。

5 おわりに

本論文では名詞間の類似度を適切に測ることを目的に、Web ディレクトリを利用して名詞のジャンルベクトルを構築した。具体的には Web ディレクトリの 2 階層目のジャンルをジャンルに設定し、3 階層目のジャンルのタイトルからジャンルの種名詞を作成し、コーパスから得た対象名詞と種名詞との関連度から、対象名詞とジャンルとの関連度を測った。名詞クラスタリングと文書クラスタリングの実験を行い、構築したジャンルベクトルが有効であることを示した。今後は種名詞を手作業で作成することで、より適切なジャンルベクトルを構築したい。

参考文献

[1] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58, 2007.

[2] 橋本力, 黒橋禎夫. 基本語ドメイン辞書の構築と未知語ドメイン推定を用いたブログ自動分類法への応用. Vol. 15, No. 5, pp. 73–97, 2008.