自己相互情報量を用いた文の接続性の判定について

鈴木 千統　　指導教員：山村 毅

1 はじめに

インターネットの発達により，私達の周りには文章が氾濫している．これらの文を要約することは、私達に検索可能な文章を増やしてくれ有益である．要約において文章の意味的なまとまりは重要な要素であるが，インターネット上の文章は意味的に適切なまとまりで分割されていない場合も多い．そこで文章を意味的に分割する技術もまた有益である言える．今回，意味的にまとまった文章を，意味的に連続した文の組の集合と考える．この時，文同士の意味的な接続性を判定できれば文章を意味的なまとまりに分割できる．

本研究では，特定の単語を含む文の次の文には共起しやすい単語が存在すると仮定し，この時の自己相互情報量 (PMI) を用いることによって 2 単語間のつながりの強さを数値化し，それを用いて 2 文間のつながりの強さを数値として出し，閾値によって接続性の判定を行う．

以降において関係を測定する 2 文の内，先に来るものを前文，後に来るものを後文と呼ぶことにする．

2 つながりの強さの計算

2.1 文のつながりの考察

まず，文のつながりについて考察する．次の 2 組の文について見てみる．

組 1：私は「夏祭り」に行った。『輪投げ』が楽しかった。

組 2：私は「夏祭り」に行った。『バスケ』が楽しかった。

ここで，2 つの組の違いは『』で囲まれた単語だけである．ここでは組 1 のほうが組 2 よりもつながりが強いと考える場合が多い．これは，「夏祭り」と『輪投げ』,『バスケ』の共起が関係していると考え，単語同士の共起が文のつながりの強さの指標に使えるのではないかと考察した．

2.2 PMI

単語同士の共起を表す指標として自己相互情報量 (PMI) がある．

P(A) を事象 A の生起確率，P(B) を事象 B の生起確率であるとし，P(A,B) を事象 A,B の同時生起確率とする．この時の単語間の自己相互情報量 (PMI) は次の式で求められる．

$$PMI(A,B) = \log \frac{P(A,B)}{P(A)P(B)}$$

ここで，次の文の組が存在すると仮定する．v_i, w_j は単語を表す．

前文：$v_1, v_2, ..., v_n$
後文：$w_1, w_2, ..., w_m$

この時，文の共起は PMI($(v_1, v_2, ...v_n)$,$(w_1, w_2, ..., w_m)$) によって求めるのが理想的ではあるが，実際の文においてこれを求める場合，必要となる学習データが莫大な量となり現実的でない．そこで PMI(v_i, w_j) を個別に求め，これを利用することによって 2 文間のつながりの強さを数値化する．

2.3 ゼロ頻度問題

また，ゼロ頻度問題解決の為に加算スムージングを行う．この時，P(X) は次の式で求められる．

$P(X) = \log \frac{n_1+s}{N_1+V_1 s}$
n_1：単語 X の出現回数
N_1：総単語数 (学習データ)
V_1：単語の種類数 (全データ)
s：重み付け

P(A,B) は次の式で求められる．

$P(A,B) = \log \frac{n_2+s}{N_2+V_2 t}$
n_2：単語の組 (A,B) の出現回数
N_2：単語の総組み合わせ数 (学習データ)
V_2：単語の組み合わせ種類数 (全データ)
t：重み付け．

3 使用データ

3.1 つながりのある文

学習データとテストデータを作成するにあたって，つながりのある文の組を定義する必要がある．今回は同一段落上の連続した文の組をつながりのある文の組として，学習データ及びテストデータに使用した．

3.2 つながりのない文

また，提案手法の精度を確認するにあたって，つながりのない文の組もテストデータとして用意する必要がある．つながりのある文の組のデータから，10 組に 1 つの割合で前文と後文を交互に取り出してコーパスを作成し，そこから 2 つづつ文を文の組としてまとめたものをつながりのない文の組とした．この作成手法の都合上，本実験におけるつながりのない文の組は明示的につながりのない文の組ではなく無作為に抽出した文の組に近い．

3.3 コーパス

毎日新聞の記事データに対し，次のルールに従い文と単語を抽出する．

・名詞，動詞，形容詞のみを用いる
・同一段落上から連続した文を文の組として選ぶ
・句点で文章を文に区切る
・「」で囲まれた文は、句点があっても区切らない
・()<>【】で囲まれた文は削除する
・『』・』以外の記号を含む文は使用しない
・ 1 つの文に含まれる同一の単語は 1 回まで数える
・「する」などのストップワードを除外する

形態素解析には形態素解析ソフト MeCab を用いた．実際の学習データとしては，2009 年度の記事から抽出した 457,630 組

を用い，テストデータとして 2008 年の記事からつながりのある文 10,000 組とつながりのない文 10,000 組を用いた．

4 接続性の判定

実際に単語同士の PMI をすべての組み合わせについて求め，それ用いて文のつながりを数値化し，閾値を用いて接続性の判定を行う．その際にいくつかの計算方法を用意した．

ここで重み s,t は [1000,100,10,1,0.1,0.01,0.001,0.0001,0.0001] の中からそれぞれ 1 つ，F 値の平均が最大になるようにして決めるものとする．また文の接続性を判定するための閾値は一定の範囲を 0.1 づつ動かし，F 値の平均が最大となったものとする．ここにおける F 値の平均とは，つながりありを真陽性 (TP) とした場合の F 値と，つながりなしを真陽性 (TP) とした場合の F 値の平均を指す．

4.1 手法

次の 3 つの手法によって単語同士の PMI から文同士のつながりを数値化する．

4.1.1 PMI の和

すべての PMI を足し合わせる．ここで，文の長さが結果に影響する場合，精度が下がると考えられる．

4.1.2 PMI の平均

すべての PMI の平均を取る．平均することによって文の長さの影響を減らす狙いがある．

4.1.3 PMI 乗の平均

ネイピア数 e の PMI 乗の平均をとる.PMI 乗とすることによって，小さな値の PMI の影響をより少なくし，大きな値の PMI の影響をより多くすることが可能となる．

4.2 結果

結果を表 1 から表 3 に示す．なお，(真) は実際に用いたテストデータがつながりありとなしのどちらに分類されるかを表し，(予想) は本実験の判定結果がつながりありとなしのどちらにどちらに分類されるかを表すものとする．また，F 値における () の中身は真陽性 (TP) をつながりありとなしのどちらにするかを表す．

4.2.1 PMI の和

表 1　PMI の和

	つながりあり (予想)	つながりなし (予想)
つながりあり (真)	6776	3224
つながりなし (真)	3577	6423

重み s	10
重み t	0.01
閾値	-20.5

正答率	0.660
F 値 (つながりあり)	0.666
F 値 (つながりなし)	0.654
2 つの F 値の平均	0.660

表 1 の結果から，この手法ではつながりのある文とない文を約 66% の精度で判別可能であることが分かる．また，閾値が負の値を取ることから，大きい値の出る PMI が少ないことが予想できる．

4.2.2 PMI の平均

表 2　PMI の平均

	つながりあり (予想)	つながりなし (予想)
つながりあり (真)	6968	3032
つながりなし (真)	3324	6676

重み s	10
重み t	0.01
閾値	-0.5

正答率	0.682
F 値 (つながりあり)	0.687
F 値 (つながりなし)	0.677
2 つの F 値の平均	0.687

表 2 の結果から，この手法ではつながりのある文とない文を約 68% の精度で判別可能であることが分かる．表 2 の 2 つの F 値の平均は，表 1 の 2 つの F 値の平均よりも大きな値としてでており，文の長さの違いが判別の精度を下げていることが分かる．

4.2.3 PMI 乗の平均

表 3　PMI 乗の平均

	つながりあり (予想)	つながりなし (予想)
つながりあり (真)	6771	3229
つながりなし (真)	2686	7314

重み s	10
重み t	0.001
閾値	2.1

正答率	0.704
F 値 (つながりあり)	0.696
F 値 (つながりなし)	0.712
2 つの F 値の平均	0.704

表 3 の 2 つの F 値の平均は，表 2 の 2 つの F 値の平均よりも大きな値としてでている．このことから，大きな値の PMI をより重視した結果のほうが良いことが分かる．

5 まとめ

単語同士の自己相互情報量 (PMI) をすべての組み合わせについて求め，それ用いて文のつながりを数値化することによって，文の組の接続性の判定を行った．毎日新聞を対象とした実験では，結果的に 7 割程度の精度を得ることに成功した．また，PMI 乗の平均の結果が良かったことから，小さい値の PMI よりも大きな値の PMI を重視した方が良い結果が得られと考えられる．今後の課題としては，さらなる精度を出すための計算手法の改善と，新聞記事以外のデータのに対しても効果があるかを調べること，2 つの文の組だけでなくその前後の文も考慮した計算の確立などが挙げられる．