２文間の単語の生起確率を用いた文の関係強度についての考察

情報科学科　鈴木 千統　　指導教員：山村 毅

1　はじめに

近年,インターネットの普及により私たちの入手可能な文章は増加の一途をたどっている.そのため１つの文章の理解に消費する時間的コストを減らすことは有益だと言える．ここで文同士の関係が分かっている場合，これを用いることによって文章をまとまりに分割し,より効率的に処理することが可能となる.

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,この時の生起確率を用いることによって２文間の関係強度を測定する.

具体的には,新聞記事を対象に上記の生起確率を共起確率とした自己相互情報量(PMI)用いて関係強度とした.

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置く.

2　関係強度の計算

P(B|A)を前文に単語Aが来るとき後文に単語Bが来る確率(生起確率)，P(B)を単語Bの出現率とする．この時の単語間の自己相互情報量(PMI)は以下の式で求められる.

$$\mathrm{PMI} = \log\big(P(B|A)\big) - \log(P(B))$$

前文の特定単語と後文の単語それぞれに対する PMI を全て求めこれらを平均する.これを特定単語と後文との関係強度とする．なお生起確率が 0 の時は PMI も 0 とする.同様にして前文の全ての単語と後文との関係強度をそれぞれ求め,これを平均したものを 2 文間の関係強度とする.

3　実験方法

今回,毎日新聞の記事データから同一段落上にある連続した２文を組として抽出して利用する．特に 2009 年度の毎日新聞の記事から抽出した約 58 万組を学習用のデータとした.

まず，毎日新聞の記事データに対し，形態素解析システム MeCab によって形態素解析を行い取得した単語群の内,名詞・動詞・形容詞のみを抜き出し，それらを EDR の辞書データに登録されている概念見出し辞書を用いて概念識別子に変換する.概念見出し辞書において複数候補がある場合はより上位に記述されていたものを適用した．次に,EDR の概念体系辞書に記載されている上位概念・下位概念を元に木構造を作成し,それを用いて使用される単語の深さを揃えた.複数の上位概念を持つものに関しては最も浅い上位概念にのみ繋がるものと仮定し,その上で複数候補がある場合は概念体系辞書においてより上位に記述されていたものを適用した.

特定の単語を含む文の出現回数と,その後文に関して生起された単語とその出現文数を数えたものを学習用データから作成し，これを生起回数データとする．このデータを元に生起確率を算出する.このデータに関しては必要試行回数の観点から一部のデータ削減を行った.

また,文の合計数と各単語の出現した文の数を記述したデータを作成し,これをもとに単語の出現確率を算出する.これらを用いて文の組ごとに関係強度の測定を行った.

4　実験結果

実験結果を以下の表にまとめる．なお 2009 年度の記事をテストデータとしたものは参考値として記載する.

表 1 は学習データに 2009 年度の新聞記事から得られた全データ約 58 万組を用いた時の関係性強度をそれぞれについて求め平均したものである．ここで，接続詞で繋がる組とは,全記事データ組から接続詞で明示的に繋がる文の組を全て抜き出したものを指し,無関係データ組とは全記事データ組から 1/200 の頻度で前文を抽出したものを合わせた 1000 組からなるデータである.

表 2 は学習データと同じものをテストデータとして用いた時に，全体の上位 90%を採用するような閾値 0.0459 を取った場合，残ったデータの割合を示す．この閾値は２つの文がつながるかどうかを判定する時の関係強度に対応するもので表２から 90%程度を正しく判定できることが分かる.

表 1　関係強度の平均

テストデータ	2008 年度	2009 年度
全記事データ組	0.2270	0.3735
接続詞で繋がる組	0.2205	0.3458
無関係データ組	−0.0346	−0.0400

表２　閾値 0.0459 以上となる割合

テストデータ	2008 年度	2009 年度
全記事データ組	0.791	0.900
接続詞で繋がる組	0.840	0.932
無関係データ組	0.305	0.310

実験結果より,無関係データ組を用いた場合の結果はそうでない場合に比べて低い値が出る傾向にある事が読み取れる.

また，ここで 2008 年度の記事における変動係数を求めると，全記事データ組の時 1.36,接続詞で繋がる組の時 0.98 となる．ここから全記事データ組を用いた結果はばらつきが大きく安定しないことが分かる.

表 2 より，文のつながりの判定において閾値 0.0459 以上となるものを採用すると,誤識別率は全記事データ組で約 20%，無関係データ組で約 30%と出る.

5　まとめと考察

実験結果より,連続した 2 文から得られた単語の生起確率を用いることによって,文同士の関係強度をある程度推測できると考えられる.

今回作成した無関係データは,そもそも学習データと同じ 2009 年の時点で約 30%の誤識別率が存在する為，これ以上の精度の向上には関係強度測定の式の改善や生起データの絞り込みが必要であり，今後の課題となる.