1ZA-6

形態素の共起確率を用いた国語教科の選択問題の解答システム

伊倉　永賢　　松井　祥峰　　乾　伸雄　　小谷　善行
東京農工大学情報コミュニケーション工学科

１．　はじめに

質問応答や情報検索の研究は重要な分野でありその研究の一つとして選択問題を解くという研究はこれまでにいくつか行われてきた．しかしこれまでの研究では選択肢が単語となっている問題が扱われていた．そこで本研究では選択肢が文章である問題を形態素の共起関係という統計的手法を利用することにより解答することを目的とする．共起関係の測定には相互情報量を使用した．

２．　国語教科の選択問題

本研究で扱う選択問題とは「本文」と本文について問う「問題文」，複数の「選択肢」から構成される．解答は問題文の意図に沿った数だけ選択肢から選択する．

３．　解答の手法

複数の選択肢から解答を求める方法として本研究では相互情報量を利用して選択肢ごとに本文との類似度を求め，類似度の高い選択肢を解答として出力する．また相互情報量に重みを掛けることにより形態素の重要度を変えて実験を行う．

３．１　相互情報量を利用した計算

選択肢から解答を選択する方法として相互情報量を利用した計算を行う．選択肢に含まれる単語を$q=\{q_1,q_2,\cdots,q_n\}$，本文内のある段落kに含まれる形態素を$m_k=\{m_{k1}m_{k2},\cdots,m_{kT_k}\}$，ある段落に含まれる形態素の総数を$T_k$と置く．$m$は$m=\{m_1,m_2,\cdots,m_k\}$とする．相互情報量とは形態素$q_i$と形態素$w_{kj}$の共起確率$P(q_i,m_{kj})$とそれぞれの形態素が個別に出現する確率$P(q_i)P(m_{kj})$の比であり，式(1)で表される．また，今回使用するコーパスは毎日新聞 1991 年データ集である [4].

$$I(q_i,m_{kj})=\frac{P(q_i,m_{kj})}{P(q_i)P(m_{kj})} \quad \cdots (1)$$

ある一つの選択肢の類似度を求める式を(2)に示す．W_{ijk}は相互情報量$I(q_i,w_{kj})$に掛け合わせる重みを表す．uは一つの選択肢に含まれる形態素の総数を，α，βは段落の範囲を示す変数である．(2)は選択肢の全ての形態素と，本文の段落を単位とした一定の範囲に含まれる形態素との相互情報量を全て加算し，平均を出力する．重みW_{ijk}については３．２で記述する．

$$sim(m_k,q)=\frac{1}{u\sum_{k=\alpha}^{\beta}T_k}\sum_{i=1}^{m}\sum_{k=\alpha}^{\beta}\sum_{j=1}^{Tk}W_{ijk}I(q_i,m_{kj}) \quad \cdots(2)$$

３．２　重み付け

式(3)で掛けている重みW_{ijk}の決定方法としてIDF 値を利用する方法，品詞を限定する方法，段落間の距離による重み付けの方法を使用する．

IDF 法の式を(3)に示す．Nをコーパスに含まれている総文書数とする．IDF 法により求めた値を重みとすることにより，重要ではない出現頻度の多い形態素を含む場合の値を小さくすることができる．

$$W_{ijk}=\log\left(\frac{N}{q_i\text{と}m_{kj}\text{が出現する文書数}}+1\right) \quad \cdots(3)$$

品詞を限定する方法は，共起関係を測定する品詞を限定して選択肢の類似度を測定する．助詞や助動詞のように出現頻度の高い品詞は他の品詞に比べて値が非常に大きくなると考えられる．そのため他の品詞同士の相互情報量が反映されにくい．相互情報量を求める際に除いた品詞を含む

An Answering System of Multiple-Choice Sentential Problems in Literacy Learning by Word Co-occurrence Relation

Norikazu IKURA, Yoshitaka MATSUI, Nobuo INUI, Yoshiyuki KOTANI,
Tokyo University of Agric. And Tech., Dept. of Computer, Information and Communication Sciences.

場合，重みW_{ijk}を 0 として掛ける．IDF 値を掛け合わせる方法と違う点は IDF 値を重みとする場合，頻出する形態素の共起頻度は小さな値となるが，この方法ではすべて 0 となる．

段落間の距離による重み付けは，問題文で問われている部分を含む段落から，段落を単位とした距離により一定の重みを相互情報量に掛ける方法である．図１に例を示す．重みの値は段落単位となり，形態素には依存しない．設問部分に近い形態素ほど設問部分の近い形態素ほど重要であるため共起関係が高いと考えた．

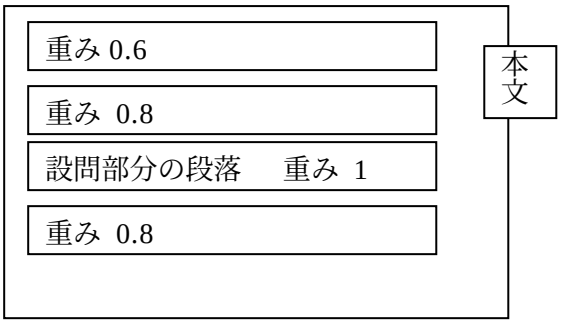

図１　段落間の距離による重みつけの例

４．　実験結果と考察

本研究の実験結果と考察を以下に示す．形態素解析は[3]を使用した．本稿で示す結果は，本文に含まれるすべての単語と共起関係を測定した場合(方法１とする)と，段落間の距離による重み付けを行った場合(方法 2 とする)の結果である．段落の重みは問題文で問われている部分を 1 とし，段落間の距離が 1 増えるごとに重みを 0.2 減算している．実験に用いた国語問題は，センター試験国語Ⅰ・Ⅱの問題である [2]．問題の例を以下に表１として示す．問題数 100 題に対して行ったシステムの実験結果を表２に示す．

表１　選択肢の部分の例

①人間の主観と客観の混合した直観の世界が、再び主観と客観に区分されること。
②我々が熱中のあまりわれを忘れた状態から目覚め、冷酷な自分を取り戻すこと。
③私の意識が、意識するものと意識されるものに分裂し、知る働きが現れてくること。
④人間の意識の根源にある世界が、見る私と見られる対象の世界に分離すること。
⑤私と私でないものの世界が、明確に分かれて意識の世界に顕在化すること。

表２　実験結果

	問題数	正解数	精度
方法 1	100	27	27%
方法 2	100	29	29%

選択肢の数は五つであるため精度はランダムに解答を選択する場合よりはよい精度となった．しかし方法１と方法２において精度にそれほど大きな差が現れなかった．その理由として，形態素の出現頻度を見ると「の」や「は」などの助詞や助動詞，句読点などを含んだ場合の共起関係の値が大きくなっており，助詞や助動詞を含まない形態素同士の共起頻度が小さくなっていた．出現頻度の多い形態素を含む共起頻度の影響が大きいため，段落間の距離による重みを掛けても共起頻度の差に影響がなかったためであると考えられる．これより IDF 値を重みとして掛けた場合と特定の品詞を除いて測定を行う場合は精度がよいと考えられる．

５．　おわりに

本研究では共起関係を利用して国語問題を解答した．結果は形態素の共起関係のみを用いたことを考えるとよい結果であった．

[参考文献]

[1] Egidio Terra，C.L.A.Clarke：Frequency Estimates for Statistical Word Similarity Measures，Proceedings of HTL/NAACL，2003

[2] 教学社出版センター：2005 年度版大学入試センター試験過去問研究 4 国語Ⅰ・Ⅱ，教学社，2004

[3] 奈良先端科学技術大学院大学自然言語処理学講座：日本語形態素解析システム「茶筅」，http://chasen.naist.jp/hiki/ChaSen/

[4] 毎日新聞社：CD-毎日新聞 '91 データ集，日外アソシエーツ