科学論文における要旨–本文間のハイパーリンク自動生成

宮川 達彦　建石 由佳　辻井 潤一
東京大学 理学部 情報科学科
{miyagawa, yucca, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

WWWの急速な普及によるオンライン文書データベースの増大にともない，膨大な文書データから自分の必要とする情報を効率よく獲得するのは，研究者にとって大きな負担となっている．これに対し，論文内全文検索エンジンのような従来の文書単位の情報検索システムは，ある情報に関連のある文書の集合を獲得することはできるものの，その結果得られた文書から情報を効率よく獲得することには利用できない．この不便を解決するため，新たに獲得した文書に対する読書支援システムとして，自動要約 (アブストラクト自動生成) [1, 2] や関連文書間のハイパーリンク生成 [3, 4] などの研究がなされてきた．

本稿では，科学論文に著者もしくは専門家がつけたアブストラクト (論文要旨) の各文から，本文の対応箇所へのハイパーリンクを自動生成し，研究者が新たな論文を獲得した際の「斜め読み (skimming)」を支援するシステムを提案する．このシステムを用いることにより，ユーザはアブストラクトを最初に読み，自分の必要とする知識が記述されている場合にその記述を論文中から瞬時に検索することが可能となる．

科学論文のアブストラクトは，本文からの抜き出しあるいは複数の文のまとめ，言い換えであることが多い [5] ことから，従来の情報検索で用いられてきた，$tf \times idf$ 法の重みづけを利用したベクトル空間モデル [6] による類似度計算が適用できる．また，アブストラクトは本文の要約としての役割も持つことから，自動要約で従来用いられてきた種々の手法をとりいれることによって，精度の向上が期待できる．本稿では，その手法のうち，語彙的連鎖およびいくつかのヒューリスティクスを導入し，実際に分子生物学論文に対して適用してその有効性を確認する．

2 システム構成

2.1 システム概要

本システムでは，WWW上のURLや，SGML形式に変換したテキストを入力として，論文内のアブストラクト–本文間の類似度を計算し，関連する箇所へのリンクが張られたハイパーテキストを出力する．

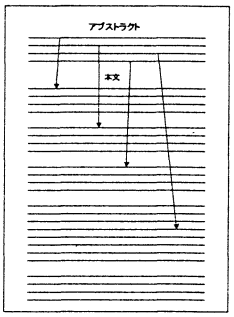

図 1: ハイパーリンクの概念図

実装された斜め読み支援ツールにおいては，入力と出力は *Netscape Navigator* などのウェブブラウザをインターフェースとして行い，読書支援に活用する (図 2)．

ハイパーリンク生成は，アブストラクトの各文と本文中の各段落との間で行う．類似度の計算には，情報検索で広く用いられてきた，$tf \times idf$ 法を基にしたベクトル空間モデル [6] を利用する．文書の談話構造を類似度計算に反映するため，語彙的連鎖 (Lexical Chain)[1] やさまざまなヒューリスティクス [5] から得られる情報を用いて $tf \times idf$ 法の単語の重みづけや，ベクトル空間モデルによって計算される類似度の補正を行う．語彙的連鎖の生成に必要な関連語辞書は，システムの前段階として論文コーパスから生成する．

2.2 関連語辞書の生成

語彙的連鎖とは，意味的なつながりを持つ単語の集まり [4] であり，文書の談話構造を表す．本稿では，このような意味的なつながりを持つ単語の集合を**意味クラス**と呼ぶ．語彙的連鎖を構成している単語は，そのセグメント (意味的なつながりを持つ文書内小単位) 内での主題を表していると考えられる．意味クラスの構成にはシソーラスが必要であるが，本システムでは分子生物学論文を対象としているため，既存の辞書には登録されていない単語が頻繁に出現する．このために，医学分野の概念辞

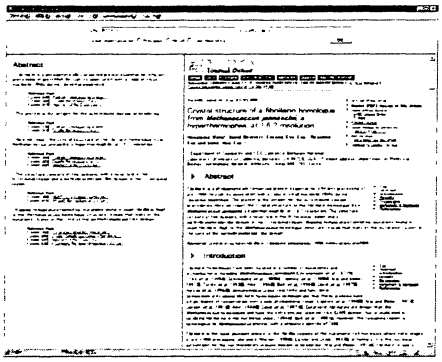

図 2: 斜め読み支援ツール

書である MeSH tree[7] から得られるシソーラスに加え，分子生物学論文コーパスから自動生成した関連語辞書を組み合わせて意味クラスを生成する．

一般に，文書中で多く共起する単語の組には意味的なつながりがあると考えられる．そこで，同一段落で共起した単語の組についてその共起しやすさを表す相互情報量 [8] を以下の式で計算する．

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

ここで，$P(x)$ は x がある段落に出現する確率，$P(x,y)$ は x,y がある段落で共起する確率を表す．本システムでは，共起回数が 5 回以下の組は不正に高い値を出すので無視する．この計算により得られた相互情報量が一定の閾値以上となる単語の組を，同一の意味クラスに属すると見なし，MeSH tree 上で親子関係にある単語の組とマージして関連語辞書を生成する．

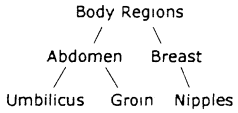

図 3: MeSH tree の木構造の例

2.3 類似度計算

新たに獲得された論文に対して，以下の手順で類似度を計算し，ハイパーリンクを生成する．

1. 論文を SGML 形式に変換し，アブストラクトおよび本文の構造を認識する．
2. アブストラクト，本文の各文中の単語をそれぞれ原型に変換する．*this, that, of, the ...* のような内容に関係がない**ストップワード**に登録されている単語は，以降の計算では無視する．
3. アブストラクトの各文と本文中の各段落との間で類似度を計算する．一定値以上の類似度を持った組み合わせについてハイパーリンクを生成する．

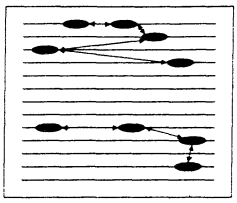

図 4: 語彙的連鎖の概念図

3 の類似度計算を詳しく述べると以下のようになる．

1. 各段落のベクトルを構成する単語は，$tf \times idf$ 法に基づいて重みをつける．ただし，本研究では同一文書内でのリンク生成を目的としているため，idf の単位を，従来定義されている**文書**ではなく，以下のように**段落**として計算する．

$$\begin{aligned} tf(p,t) &= \text{単語 } t \text{ が 段落 } p \text{ 中に出現する頻度} \\ idf(t) &= \log \frac{\text{すべての段落の数}}{\text{単語 } t \text{ が出現する段落の数}} \end{aligned}$$

2. 2.2 節で生成された関連語辞書から語彙的連鎖を構成する．具体的には，本システムは図 4 のように同一の意味クラスに属する単語が連続して現れる場合を検出し，語彙的連鎖を構成して出力する．連鎖の構成の際には，連鎖の長さと各連鎖間の距離に関して 3 つの閾値を必要とする．本システムでは文書長などを考慮し，最大連鎖長：*500* 語 / 最小連鎖長：*100* 語 / 空白閾値：*30* 語 に設定した．

   実際に連鎖を構成し，この手法が適切な語彙的連鎖を構成しているかを詳細に調べてみたところ，そのセグメント内での主題となる単語についての語彙的連鎖は正しく構成できているが，それに加え，文書全体の主題を表す単語が，文書中のいたるところで適切でない語彙的連鎖を構成してしまうことがわかった．これは，「文書全体の主題を表す単語は論文中にまんべんなく出現するため，そのセグメントでの主題でなくても，語彙的連鎖を構成してしまう」ことが原因であると考えられる．

   そこで本システムでは，同一の意味クラスを表す語彙的連鎖が同一セクションで複数構成された場合，最初に出現したもののみを有効な連鎖として扱う．

3. 語彙的連鎖 C を構成している単語の tf 値を補正するため，連鎖の強さ $S(C)$ を以下の式で計算する [4].

$$S(C) = (C \text{ に含まれる単語ののべ総数}) - (C \text{ に含まれる異なる単語の数})$$

各段落のベクトルに含まれる単語のうち，連鎖 C を構成している単語は，その重みを $1 + L * S(C)$ 倍に補正する．L の最適値は実験により決定する．

4. ヒューリスティクスを用いて補正する．本システムに適用できるヒューリスティクスとしては，以下のものが考えられる．

   - アブストラクトに慣用表現 (*This paper describes ..., The results show ...* など) が現れた場合，文書の構造を表す特定のセクション (*Introduction, Results* など) への関連が大きい [5].
   - 段落に見出し語がついている場合は，その見出し語が段落の主題を表している．

   本システムでは，前者は対応するセクション中にある段落すべてに重み係数 p を設定し，類似度計算の際に補正する．今回は p の値は 3 とした．後者は見出し語に含まれる単語の重みに 1.5 を乗じて補正する．ただし，これらのヒューリスティクスを実際に使用するかはユーザが選択できるようにする．

5. アブストラクトの各文をクエリーとして，各段落のベクトルとの類似度をベクトル同士の *cosine* 値で計算する．クエリーベクトルは，アブストラクトの文に含まれる各単語の重みをすべて 1 としたベクトルとする．4 で重み係数 p が設定された段落については，計算された *cosine* 値に重み係数を乗じる．

3 システムの評価

3.1 実験

実験は，分子生物学 EMBO Journal Online[9] に収録されている論文を対象として行った．1000 件の論文をトレーニングコーパスとして，相互情報量による関連語辞書を構築し，新たに獲得した論文について，分子生物学分野の専門家が手動で振ったアブストラクト-本文間のリンクと，本システムによって生成されるリンクを比較して本システムで用いられる各手法の有効性を評価する．評価は，システムが出力したリンク集合 (O とする) を，専門家が手動で振ったリンク集合 (A とする) と照合して，以下により定義される，適合率 (Precision)，再現率 (Recall) を基準として行う．

$$\text{適合率} = \frac{|O \cap A|}{|O|}, \text{再現率} = \frac{|O \cap A|}{|A|}$$

ここでは，以下の 2 つの実験により語彙的連鎖とヒューリスティクスの有効性を確認する．

1. 語彙的連鎖による補正の効果を確認するため，L の値を 0 から 1 の間で 0.05 単位で動かし，ヒューリスティクスは利用せずに適合率-再現率を計算する．

2. ヒューリスティクスの効果を確認するため，1 でよい結果を示した L に対し，ヒューリスティクスで単語の重みづけおよび *cosine* 値を補正し，補正しない場合との間で適合率-再現率を比較する．

3.2 実験結果

実験には，専門家に正解となるリンクを振ってもらった 4 件の論文を使用し，本システムの有効性を評価した．

まず，語彙的連鎖の補正係数 L を変化させて再現率-適合率を計測した．図 5 に，4 件のうち 1 件の論文について $L = \{0, 0.2, 0.4\}$ とした場合のグラフを示す．他の 3 件についても同様の結果が得られた．

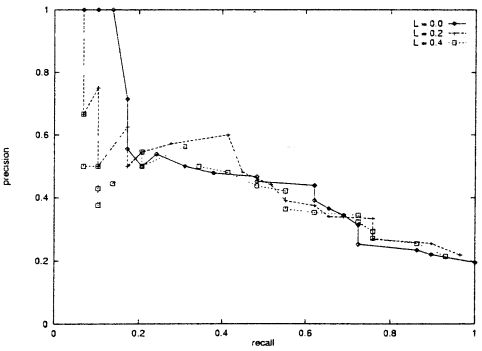

図 5: L の値による再現率-適合率の変化

図 5 から，適合率が高域に近づくと，再現率が低くなるが，逆に 再現率を 1 に近づける際に，tf 値単独の場合 (図の $L = 0.0$ の場合に相当) と比べて適合率が下がりにくい結果が読み取れる．

これは，ベクトル空間モデルでは不可能な談話構造の抽出を行い，再現率を高めるという，語彙的連鎖の効果によるものと考えられる．本システムの目的である斜め読み支援ツールでの利用を考えると，再現率が高い方が，関連箇所を逃さず検索することができて便利であるという点から，語彙的連鎖による補正は本システムにおいて有効であろう．

次に，もっとも良い結果を出している $L = 0.2$ の場合に，ヒューリスティクスによる補正を行って再現率-適合率を計算し，補正しない場合と比較した．4 件の論文のうち，3 件では再現率・適合率が中域において，補正を行わない場合より精度が向上した (図 6)．

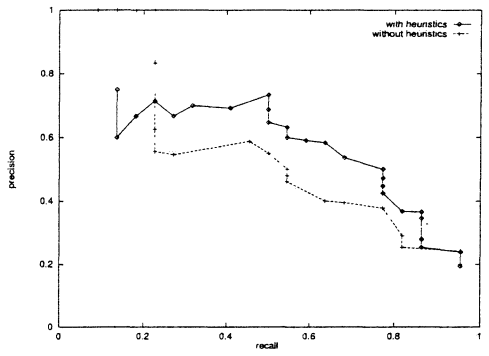

図 6: ヒューリスティクスによる補正

4 おわりに

小規模な実験データではあるが，語彙的連鎖による単語の重みづけの補正が，再現率高域での適合率の向上にある程度有効であることが確認された．ヒューリスティクスに関しては，さらに大規模な実験を行い，どのように働いているのかを解析する必要があろう．

今回の実験で，予想したほど精度が向上しなかったのは，以下が原因として考えられる．

1. 分子生物学用語では表記の揺れが多く起り，本システムで生成した関連語辞書による語彙的連鎖だけでは処理できない，
2. 本システムではアブストラクトを検索のクエリーとしており，それぞれのクエリーに含まれる単語数は 20...40 程度となるので，ベクトルが短すぎる．
3. 各文に対し，段落単位でリンクを生成しており，かならずしも段落の区切りが意味の区切りに対応しない場合がある．

これらの問題に対して，1. は固有表現抽出 (Named Entity Extraction)，2. は検索要求拡張 (Query Expansion)，3. は表層上の構造に依存しないパッセージ抽出 [10] を利用することで，精度向上に結び付けたいと考えている．

謝辞

実験の正解データを作成していただいた大田朋子博士に感謝いたします．

参考文献

[1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *ACL Workshop on Intelligent Scalable Text Summarization.*

[2] Singhal A. and Salton G. Automatic text browsing using vector space model. In *Proceedings of the Dual-Use Technologies and Applications Conference*, pages 318–324, 1995.

[3] 田中 俊一, 岡村 潤, 森 辰則, 中川 裕志. 複数関連文書間の読書支援のための類似度計算手法. 言語処理学会第 5 回年次大会 発表論文集.

[4] S.J. Green. Using lexical chains to build hypertext links in newspaper articles. In *AAAI Workshop on Knwoledge Discovery in Database*, 1996.

[5] H. Saggion. Where does information come from? In *RIFRA*, 1998.

[6] G. Salton, A. Wong, and C. Yang. A vector space model for information retrieval. *Communications of the ACM*, 18:613–620, November 1975.

[7] National Library of Medicine. Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html

[8] K. W. Church. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), November 1990.

[9] The European Molecular Biology Organization. EMBO journal online. http://www.emboj.org/

[10] 望月 源, 岩山 真, 奥村 学. 語彙的連鎖に基づくパッセージ検索. 情報処理学会自然言語処理研究会報告, pp. 39-46, 1998. 127-6.