段落見出しの自動生成に向けて

川口 人士 †　　佐藤 理史 ‡　　駒谷 和範 ‡

† 名古屋大学 工学部 電気電子・情報工学科　‡ 名古屋大学大学院 工学研究科

{hitosi_k, ssato, komatani}@nuee.nagoya-u.ac.jp

1　はじめに

情報伝達を主目的としたテキストには、なにかしらの「見出し」が付与されるのが普通である。たとえば、教科書などの書籍では、章や節に見出しが付与され、目次において、それらの一覧が提示される。これらの見出しは、求める情報が書いてある場所の発見を手助けすると同時に、そこに書いてある内容の把握を手助けする。このように、「見出し」は、情報発見と内容把握の両方の側面で、人間を効果的に支援する。

見出しは通常、章や節にのみ付与されており、段落には付与されていない。しかしながら、「段落にも見出しがあると便利である」。この考えが、本研究の出発点となっている。

本研究では、情報伝達を主目的としたテキストを対象に、各段落に見出しを自動的に付与する方法を検討する。まず 2 節では、見出しの調査を行う。3 節では、見出しに使用する段落キーワードを抽出する方法について説明する。4 節では、実際の段落から段落キーワードを抽出する実験について説明する。

2　見出しの調査

ここでは、実際の書籍の章や節に、どのような見出しが付けられているかを調査した。調査対象には、情報検索の教科書『情報検索と言語処理』[1] を用いた。

2.1　フレーズパターン

調査対象の目次から、見出し 39 件をすべて抜き出し、その表層構造を調べた。具体的には、各見出しに次のような一般化を適用し、フレーズパターンを作成した。

1. 見出しを（人手で）文節に分割する。
2. 各文節の内容語部分を変数化する。（付属語部分は、そのまま残す。）
   - 体言を中心とした文節の場合は、その内容語部分を、A、B、C のように変数化する。
   - 用言を中心とした文節の場合は、その内容語部分が動詞の場合は V に、形容詞の場合は K に変数化する。

このような一般化を行なった結果を表 1 に示す。さらに、それらをタイプ別に集計したものを、表 2 に示す。

表 1: 見出しとフレーズパターン

第 1 章	情報検索とは	A とは
1.1	情報の蓄積と利用	A の B と C
1.2	情報検索へのアプローチ	A への B
1.3	情報検索システムの評価基準	A の B
1.4	文献案内	A
第 2 章	情報検索の基礎	A の B
2.1	文書とその表現	A と B
2.2	索引付け	A
2.3	検索質問の表現	A の B
2.4	検索質問拡張	A
2.5	検索モデル	A
2.6	文献案内	A
第 3 章	情報検索システムの性能評価	A の B
3.1	システムの性能評価の観点	A の B の C
3.2	システムの有効性	A の B
3.3	再現率と精度	A と B
3.4	その他の評価尺度	A の B
3.5	テスト・コレクション	A
3.6	新しい評価手法	K+A
3.7	文献案内	A
第 4 章	言語処理技術の利用	A の B
4.1	言語処理の概要	A の B
4.2	言語処理を利用した索引付け	A を V+B
4.3	検索質問拡張とシソーラス	A と B
4.4	文献案内	A
第 5 章	ユーザ・インタラクション	A
5.1	情報検索システムとユーザ	A と B
5.2	適合性フィードバック	A
5.3	対話による検索	A による B
5.4	進化的探索	A
5.5	エキスパート・システム技術の利用	A の B
5.6	情報検索における ユーザ・インタラクションの設計	A における B の C
5.7	文献案内	A
第 6 章	情報検索の関連技術	A の B
6.1	情報抽出	A
6.2	テキストの自動要約	A の B
6.3	テキストの自動分類	A の B
6.4	情報フィルタリング	A
6.5	文献案内	A

表 2: フレーズパターンの分類

	A	A の B	A と B	その他	計
章	1	4	0	1	6
節	14	8	4	7	33
計	15	12	4	8	39
	(38 %)	(31 %)	(10 %)	(21 %)	

誘電体とは、物理的には伝導電子を持たない絶縁体的電子構造を持ち、電界を加えると誘電分極を発生する固体、液体、気体物質の総称である。誘電体は電気・電子工学分野での材料として導体、半導体、磁性体などと並んで重要な地位を占め、主として誘電率が大きいことを利用したコンデンサ材料と絶縁抵抗の高いことを利用した電気絶縁物質材料として用いられる。

図 1: 1 段落のテキスト (出典:[2])

表 2 に示すように、タイプ「A」と「A の B」の見出しが、全体の 69 %を占める。これらのタイプの見出しは、他の教科書や書籍でも頻繁に用いられていると考えられる。本研究では、これら 2 つのタイプの見出しに焦点を当てる。

2.2 見出しの構成要素

次に、見出しの構成要素である A（や B）に、どのような内容語が用いられているかを調べた。テキストの内容を端的に表すという見出しの役割上、A として用いられる語は、そのテキストの重要なキーワードであり、テキスト中に出現している可能性が高いと考えられる。一方、B は、A を補足する語であり、かならずしもテキスト中に存在しないことが予想される。ここでは、まず、A として用いられる語がテキスト中に存在するかどうかを調べた。

タイプ「A」では、各章の最後の節の見出し「文献案内」を除き、A として用いられる語は全てテキスト中に出現していた。タイプ「A の B」では、1 つの例外を除き、すべてテキスト中に出現していた。この例外は「エキスパート・システム技術」(5.5 節の A）であり、テキスト中には「エキスパート・システムの技術」という形で出現していた。

3 段落キーワードの抽出法

前節の調査に基づき、与えられた段落から、見出し「A」「A の B」の A にふさわしいキーワードを抽出する方法を検討する。ここでの目標は、たとえば、図 1 に示す 1 段落のテキストを入力として、見出し「誘電体の定義」の作成に必要なキーワード「誘電体」を出力とすることである。

テキストからキーワードを抽出する代表的な方法に、TF-IDF 法 [Salton 89] がある。比較的長いテキストに対しては、頻度情報を利用する TF-IDF 法は有効に機能することがよく知られているが、本研究が対象とするテキストは、1 段落と非常に短いため、頻度情報だけでは不十分だと考えられる。そこで、頻度情報に加え、位置情報と文脈情報を利用することを考える。

具体的には、テキスト T に出現するキーワード候補 w に対して、次式で定義されるスコアを計算し、その値が最大となるものを選択する。

$$score(w,T) = \ _f s_f(w,T) + \ _l s_l(w,T) + \ _c s_c(w,T) \quad (1)$$

ここで、s_f、s_l、s_c は、それぞれ、頻度情報、位置情報、文脈情報に基づくスコアであり、 $_f$、 $_l$、 $_c$ は、それらの重みである。以下では、これらのスコアの計算法について述べる。なお、これらのスコアの計算では、形態素解析システム JUMAN と構文・格解析システム KNP を利用する。

3.1 キーワード候補

与えられたテキスト T の各文を形態素・文節解析し、テキスト T に出現するすべての文節を求める。これらの文節リストから体言文節を取り出し、その内容語部 (の代表表記) をキーワード候補 w とする。キーワード候補 w は、一般に、複数の構成要素 (名詞) から構成される。これらの構成要素を c_i と表す。すなわち、キーワード候補 w は、次のように表される。

$$w = c_1 c_2 c_3 \quad c_n \quad (2)$$

ここで、n は、キーワード候補 w を構成する要素の数を表す。

3.2 頻度情報スコア

テキスト T におけるキーワード候補 w の頻度スコア $s_f(w,T)$ を、以下の式で計算する。

$$s_f(w,T) = \text{TF-IDF}(w,T) + B(w,T) \quad (3)$$

第 1 項の TF-IDF(w,T) は、以下の式で計算する。

$$\text{TF-IDF}(w,T) = \text{TF}(w,T)\ \text{IDF}(w) \quad (4)$$

$$\text{IDF}(w) = \frac{1}{\log_{10}(\text{DF}(w,D)+10)} \quad (5)$$

ここで、TF(w,T) は、テキスト T における語 w の頻度を表し、DF(w,D) は、ある文書集合 D(実験では、BCCWJ の一部を用いた) において w が出現する文書数を表す。式 (5) の分母で 10 を足してから log を計算するのは、DF の値が 0 の場合も IDF が値を持つようにするためである。

式 (3) の第 2 項 $B(w,T)$ は、w が複数の構成要素を持つ場合のボーナスである。たとえば、キーワード候補 w が「誘電体」の場合、この語は、「誘電 (c_1)」と「体 (c_2)」の 2 つの名詞から構成される。図 1 に示した段落では、「誘電分極」や「誘電率」などの「誘電」を含む語が複数回出現している。このことは、「誘

電」に関する語が重要であることを示唆する。一方、「体」のような一般的な語（DF(w) の値が大きいもの）は、出現回数が多くてもそれほど重要ではないと考えられる。これらのことを考慮し、$B(w,T)$ を次式で計算する。

$$\mathrm{B}(w,T) = \frac{\sum_{i=1}^{n} (\mathrm{TF}(c_i,T) - \mathrm{TF}(w,T))\,\mathrm{IDF}(c_i)}{n} \quad (6)$$

この式は、語 w の構成要素 c_i がそれ以外の場所で出現する回数 $(\mathrm{TF}(c_i,T) - \mathrm{TF}(w,T))$ に、$\mathrm{IDF}(c_i)$ を掛け、それらの平均値を求めることを表す。なお、語 w が 1 語 $(n=1)$ の場合は、$B(w,T)=0$ となる。

3.3 位置情報スコア

位置情報スコアは、キーワード候補 w が段落 T 中のどこに出現するかに基づいて計算する。

段落中の文の重要度は、最初の文がもっとも高く、次に最後の文が高いと考えられる。テクニカルライティングでは、各段落の先頭に、その段落で最も重要な内容を記述する文（トピックセンテンス）を置くことを基本とする。これに加え、日本語では、段落の最後にまとめの文を置く書き方もしばしば見受けられる。文中の語の重要度もほぼ同じ傾向を示し、一般に、文頭付近に現れる語 (主題であることが多い) が最も重要で、その次が文末付近に現れる語である。

このような位置に基づく重要度を数値化するために、キーワード候補語 w のそれぞれの出現に対して、位置 p を割り当てる。この位置 p から、次の 4 つの値を計算できるものとする。

1. fs(p): 段落内の文番号 (正順)
2. rs(p): 段落内の文番号 (逆順; 後から何番目か)
3. fb(p): 文内の文節番号 (正順)
4. rb(p): 文内の文節番号 (逆順)

次に、あるリストのある位置 x の重要度を計算する関数 $\mathrm{z}(x,y)$ を次式で定義する。

$$\mathrm{z}(x,y) = \frac{1}{x^2} + \beta\,\frac{1}{y^2} \quad (7)$$

ここで、x はそのリストにおいて前から何番目かを、y は後から何番目を表す。すなわち、リストの長さは $x+y-1$ となる。和を計算する際の第 2 項の β は前後の重みを表し、本研究では、$\beta=0.8$ を用いる。たとえば、長さ 6 のリストの場合、式 (7) の値は、図 2 のようになる。

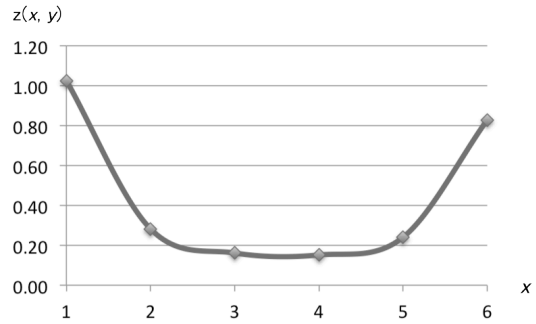

図 2: 関数 $\mathrm{z}(x,y)$

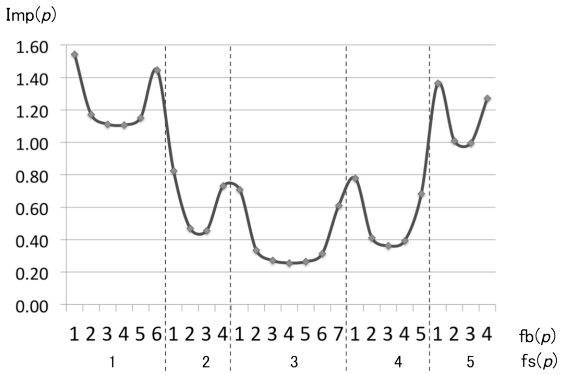

図 3: 位置の重要度 $\mathrm{imp}(p)$

以上の準備に基づき、段落 T 内のある位置 p の重要度を次のように定義する。

$$\mathrm{imp}(p) = z(fs(p), rs(p)) + \gamma\; z(fb(p), rb(p)) \quad (8)$$

ここで、γ は重みである（実験では、$\gamma=0.5$ を用いた）。段落内の位置の重要度の例を図 3 に示す。

最終的に、位置情報スコア $s_l(w,T)$ を次のように計算する。

$$s_l(w,T) = \max_{p \in loc(w,T)} \mathrm{imp}(p) \quad (9)$$

ここで、$loc(w,T)$ は、キーワード候補 w のすべての出現位置を返す関数である。すなわち、そのキーワード候補 w が出現した位置のうち、最も重要な位置のスコアを採用する。

3.4 文脈情報スコア

文脈情報スコアは、キーワード候補 w の直前または直後に手がかり語が現れた場合のボーナスから計算する。与えるボーナスの一覧を表 3 に示す。なお、助詞のボーナスの設定には、文献 [3] を参考にした。

以下の例文を用いて、ボーナスの計算法を説明する。

例 ：日本は経済大国である。ゆえに外国人労働者が多い。

表 3: 文脈情報のボーナス

文脈情報		加点
直後に助詞	は	1.2
	が	1.0
	を	0.8
	に	0.7
	の	0.6
	こそ	0.5
	で	0.4
	と、も	0.3
	へ、まで、より	0.2
	その他	0.1
直後に判定詞		0.5
直前に接続詞		0.5

この例では、直後に助詞が現れる「日本」と「外国人労働者」、直後に判定詞である「経済大国」、および、直前に接続詞が現れる「外国人労働者」にボーナスを与える。この結果、たとえば、「外国人労働者」のボーナスは $1.0 + 0.5 = 1.5$ となる。

文脈情報のスコア $s_c(w, T)$ は、キーワード候補 w のすべての出現位置における、これらのボーナスの最大値を採用する。

4 実験

3 節で述べた段落キーワード抽出法を、実際の段落に適用する実験を行なった。

4.1 準備

実験の準備として、情報伝達を主目的とするテキストを複数用意し、その中から 50 段落を選び、人手で理想的な見出しを「A の B」の形で付与した。ただし、「A」は、かならず段落中にある語から選んだ。なお、このデータは、システムの開発でも参照した。

頻度情報の $DF(w, D)$ の値の計算に使用する文書集合 D には、「現代日本語書き言葉均衡コーパス (BCCWJ)」を用いた。具体的には、$DF(w, D)$ の値を求めるために、BCCWJ の長単位の表形式の形態論データ (LUW) を使用した。使用した BCCWJ のサンプル数（文書数）は約 17 万件、形態素数（長単位）は約 1000 万件である。

4.2 実験結果

図 1 に示した段落にキーワード抽出法を適用し、頻度情報スコア、位置情報スコア、文脈情報スコアを求めた結果を表 4 に示す。この表のスコアの合計は、$_f = _l = _c = 1$ で計算してある。この例では、キーワード候補の第 1 位として、「誘電体」が得られた。

次に、$_f$、$_l$、$_c$ の値の重みについて検討した。今回は、$_l = _c = 1$ とし、$_f$ の値を 0.1〜1.0 まで 0.1 刻みで変更したところ、$_f = 0.4$ のときに最もよい結果が得られた。そのときの、50 段落に対する段落キーワード抽出結果を表 5 に示す。この表では、上位 N 位までに、人手で付与した正解と一致するものが含まれるどうかを示している。今回作成した 50 段落に対しては、半分の 25 段落に対して、正解キーワードを抽出することができた。表 5 には、頻度情報スコアのみを使用した場合（$_l = _c = 0$）の実験結果を併せて示した。この表より、位置情報スコアと文脈情報スコアの導入により、正解数が向上していることがわかる。

表 4: 出力結果

キーワード候補	頻度	位置	文脈	計
誘電体	2.87	1.70	1.50	6.07
誘電率	2.19	1.06	1.00	4.24
誘電分極	1.85	1.22	0.80	3.87
絶縁体的電子構造	1.72	1.22	0.80	3.74
絶縁抵抗	1.81	1.06	0.60	3.47
伝導電子	1.37	1.26	0.80	3.43
:	:	:	:	:

表 5: 実験結果

上位 N 位	本手法		頻度のみを使用	
	正解数	割合	正解数	割合
1	25	50 %	16	32 %
3	40	80 %	32	64 %
5	42	84 %	38	76 %
10	48	96 %	43	86 %

5 おわりに

本論文では、段落見出しの自動生成に向けて、段落中のキーワードを抽出する方法を示した。今後は、大規模なテストセットを作成して、抽出法を評価するとともに、「A の B」見出しの B を抽出（または作成）する方法を検討する必要がある。

謝辞 本研究では、「現代日本語書き言葉均衡コーパス」の一部を利用した。

参考文献

[1] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.

[2] 犬石嘉雄. 誘電体現象論. 電気学会, 1973.

[3] 益岡隆志, 田窪行則. 基礎日本語文法–改定版–. くろしお出版, 1992.