1X-3

LCSに着目した英文科学技術二次文献からのキーワード抽出

藤原飛日　　石野明　　竹田正幸　　松尾文碩

九州大学大学院システム情報科学府情報理学専攻

1. はじめに

本稿では辞書を用いずにキーワード抽出のする方法として, 文章同士の最長共通部分列 (*longest common subsequence*;LCS) を求めることでキーワードを抽出する方法とその結果についての考察を述べる.

キーワードの抽出は大量の文書の管理をするために重要である. キーワードをうまく選んでやることで文書の分類, 検索が効率良くできるようになる. 例えばシソーラスはこのような目的のために用意されている. しかしシソーラスは一般的に手作業で作られているが大量の電子文書を対象とした場合手作業でのキーワードの抽出には限界がある. そのためにキーワードを自動的に文書中から抽出する必要がある.

従来からあるキーワード抽出の手法として, 直前に現れた単語から次に出現する可能性のある単語を調べる *n-gram* と呼ばれる手法がある. これは連続した単語の並びのうち, 出現頻度の高いものをキーワードとみなす手法であり, よく使われる言い回しや学術用語を抽出することが可能である. しかし実際の文章では同じ意味をもつキーワードでも格, 活用の違いや修飾語の有無により異なるキーワードとして扱われたり, 個々の単語列の出現頻度の低下によりキーワードとして扱われなくなるという問題がある. この問題を解決するために単語毎に重み付を行なったり, 文章に *stemming* を施すといった改良がなされるが, 一般にこれらの処理は, 辞書を用いて行なわれる. しかしこのような辞書は科学技術文献のように新たな単語が次々と生み出されるものを対象とした場合不適切である.

本稿で取り上げるLCSは不連続な単語列をキーワードとして扱うが, *n-gram* は連続した単語列をキーワードとして扱うという点で大きく異なる. 実際に *n-gram* を用いてキーワード抽出を行なう手法と, LCSを求めてキーワードを抽出する手法では, 抽出されたキーワードの数, 性質ともに大きく異なる.

2. LCSについて

LCSとは, 2つの文字列 $A = a_1a_2\cdots a_n$ と $B = b_1b_2\cdots b_m$ の共通の部分列のうち, もっとも長いもののことである. 例えば $A = \mathrm{abcdea}, B = \mathrm{abcade}$ とすると, $LCS(A,B) = \{\mathrm{abcde}\}$ となる. 近年LCSを求める問題は, 遺伝子解析の分野でよく取り上げられており, 様々なアルゴリズムが提案されている [2]. これらの手法では複数のLCSのうちの1つが結果として得られる.

本稿では文字を単語に, 文字列を文章にみなしてLCSを求めることで, 文章中からのキーワード抽出を行なう. *n-gram* で得られる結果は連続した単語列が現れるが,LCSで得られた結果は不連続な単語列となる. 例えば次のような2文

```
THIS IS WIRE.
THIS IS VERY LONG AND THICK WIRE.
```

が与えられた場合,*n-gram* を用いた場合 `THIS IS` が得られるがLCSの場合 `THIS IS LONG WIRE` が得られる.

3. 実験

英文科学技術二次文献データベースINSPECのクラスC(INSPEC-C)の1999年分のうち無作為に選んだ1,000件を対象にLCSを求めてキーワードを抽出する. このデータベースには他にも自由索引句 (free indexing term) と呼ばれる論文の著者自身が設定した複数のキーワードや, 統制句 (thsaurus term) と呼ばれるINSPECが設定

Extracting Keywords from Technical Scientisfic Documents by Using LCS

Asuka FUJIWARA, Akira ISHINO, Masayuki TAKEDA, Fumihiro MATSUO
Department of Informatics, Kyushu University, 1-10-6 Hakozaki, Higashi-ku, Fukuoka-city 812-8581, Japan

表 1: 実験で得られたキーワードの数

文書数	文章数	有効単語数	キーワード数	キーワードの平均単語数
1,000	10,341	58,137	29,886	2.15

表 2: 得られたキーワードの分類結果と例

比較対象	a	b	c	d	e	全体
FIT	10,336	15,249	2,243	1,086	972	29,886
TT	20,681	7,948	257	717	283	

FIT : 自由索引句
TT : 統制句

タイトル	Relationship between the ambiguity function coordinate transformations and the fractional Fourier transform	
種類	得られた結果	比較対象 (FIT)
b	ANALYSIS PRESENTED	TIME FREQUENCY ANALYSIS
b	APPLICATIONS ANALYSIS	TIME FREQUENCY ANALYSIS
a	APPLICATIONS PRESENTED	
b	APPLICATIONS TRANSFORM	FRACTIONAL FOURIER TRANSFORM
b	CASE FUNCTION	AMBIGUITY FUNCTION
a	CASE INTRODUCED	
c	FOURIER TRANSFORM	FRACTIONAL FOURIER TRANSFORM
c	FREQUENCY ANALYSIS	TIME FREQUENCY ANALYSIS
b	FREQUENCY PRESENTED	TIME FREQUENCY ANALYSIS

したキーワードのうち, 内容に即したものを編集者が一定の条件に従って 3 ~ 8 個列挙したものが記述されており得られた結果の検証が可能となっている.

3.1. 実験方法

まず文献データの抄録文を文章単位に切り分けさらにそれを単語単位に切り分け単語列を得る. 文章の区切りを',␣' と'.␣' とし, 単語の区切りをアルファベット以外の文字とした.

次に得られた単語列のうち品詞情報から得たストップワードを取り除く [1]. これは意味のない LCS が現れるのを防ぐためである. このようにして得られた文献データの中から異なる文献の抄録文 A, B を取りだし, A のすべての文章と B のすべての文章の LCS を求め得られた LCS をキーワードとみなす.

実験には Compaq AlphaServer DS20 (Alpha 21264 500MHz, Memory 4GB, Tru64 v4.0F) を使い, プログラムは Java で記述した. 計算時間はおよそ 4 日だった. 得られたキーワードの数などを表 1 に示す.

3.2. 結果の検証

まず得られた結果がどのくらい正しいのかを評価した. 評価の方法は得られた結果と INSPEC データベースに含まれる自由索引句, 統制句を比較し一致度を次のいずれかに分類した. 結果は表 2 のようになった.

a. 完全に不一致.
b. 一部の単語が一致.
c. 得られたキーワードが比較対象の一部.
d. 比較対象が得られたキーワードの一部.
e. 完全に一致

結果を見ると得られたキーワードのほとんどが自由索引句および統制句と一致していない. また, 一部一致の単語も頻度の高い単語が一致しているだけであった. しかし c,d,e の和も平均すると 1 文献当たり 4 個出現していることになりこれらのみをうまく抽出できれば文献のキーワードとして十分有益であるといえる.

表 3: キーワードの単語間の平均単語数

ストップワード	対象	全体	a	b	c	d	e
含む	FIT	4.67	4.48	5.30	2.01	5.12	3.64
	TT		4.60	4.94	1.97	4.95	2.17
含まない	FIT	2.10	1.94	2.74	1.54	2.17	1.83
	TT		2.02	2.94	1.50	2.83	1.50

そこで a,b と c,d,e ではどのような違いがあるかを調べるために, 抽出されたキーワードが文章中では単語間にいくつの単語が入っているかを調べた. 一般的に重要なキーワードというのは句として表現されていることが多い. したがって a,b に比べて c,d,e は単語間の単語数が少ないことが予想される. 結果は表 3 のようになった. 0 の場合 (連続した単語列) は *n-gram* でも抽出が可能なキーワードだが, それ以外の場合は抽出できない. しかし結果は e のようなものでもキーワードの単語間に平均して約 2 個の単語が入っていることが確認され, 本稿の手法が有効であることがわかる. さらに, 文章からストップワードを除いた場合どのようになるかも調べた. 注目すべき点は a の欄である. ストップワードを除かない文章では, キーワードの単語間の単語数が平均で 4 以上あったが, ストップワードを取り除くと, 元の 1/2 以下になる. b,c,d,e が, 元の 1/2 よりやや多いのに対して大きくことなる. したがってストップワードの多い文章にはキーワードとなりうる単語が入っていない可能性が高いことが予想される. ただし例外もあるうえに, 文章の区切りが必ずしも正しくはないのでこの予想は当たっているとは限らない.

4. 考察とまとめ

実験の結果 LCS を求めることで得られるキーワードは *n-gram* を用いて得られるそれよりも多くのキーワードを抽出できることがわかった. またキーワードの単語間に入っている単語の数を調べることでより自由索引句や統制句と一致するキーワードを抽出すること可能性がある. これは今後詳しく調べる予定である.

参考文献

[1] 西村利浩. 英文科学技術文献用機能語辞書の作成と評価. 九州大学電子工学専攻修士論文, 1989.

[2] C. Rick. *A New Flexible Alogithm for the Longest Common Subsequence Problem. CPM95*, 973:340–351, 1995.