未踏テキスト用シソーラスの自動構築システムの開発

山本英子[†] 梅村恭司[†] 舟宝貴志[†]
鈴木健二[†] 真田亜希子[†]
Chakma Junan[†] 武田善行[†]

本研究では，辞書に記載されていない単語を含む，整備されていないテキストを未踏テキストと考え，そのような未踏テキストを理解し検索するのに役立つシソーラスを辞書を用いずに，自動的に構築することを試みた．本論文では，関連語を同じように使用される単語と定義し，シソーラスは関連語の組として抽出した．これを実現するために，まず統計処理で候補となる単語を切り出すという処理を行い，その単語の二つに対して，同じように使用されているかどうかの判定を行った．本論文では，候補となる二つの単語が同じように使用されるかどうかの判定に用いた定義式と，判定された単語についての分析を報告する．

Automatic Building System Thesaurus for Brandnew Text Data

Eiko Yamamoto,[†] Kyoji Umemura,[†] Takashi Funatomi,[†]
Kenji Suzuki,[†] Akiko Sanada,[†] Chakma Junan[†]
and Yoshiyuki Takeda[†]

In this study, we try to build a thesaurus for brandnew text data which is useful to understand or retrieve the text data. Our system does not use any dictionaries. In this paper, we show a method to build the thesaurus, and analyze an experimentation result.

1. はじめに

日々増え続ける新聞記事やWWWのテキスト情報には新しい概念を示す単語が日々生成されている．この生成される単語は新規語であるが故に，ほとんど辞書に記載されていない．我々はこのような辞書に記載されていない単語が含まれるテキストは人間によって整備されていない未踏テキストと考える．未踏テキストは最新情報を記述した文書であることが多い．その最新情報を理解することやその情報に関連するテキストを検索するには，テキストに出現する新しい概念を表す単語を理解する必要がある．しかし，このような新しく生成された未知語をそのまま理解し検索に用いることは難しい．また既知語であっても，テキストが扱う分野によってはその単語の意味が他の分野で用いられる既知の意味と異なる場合や，作成者によっては同じ意味を表す単語であるのに異なる表記が用いられる場合がある．これらの場合もその単語を理解し検索に用いることは難しい．このような単語を理解するために，その単語に関する情報として関連語が考えられる．たとえば，未知語に関連する既知語を知ることができれば，その未知語を理解することができ，未知語に関連する情報を検索することができる．また，未知語に関連する他の未知語を特定することにも役立つ．そこで，本研究では，未踏テキストから未知語，既知語を問わずその単語と関連語が対となった関連語リストを未踏テキスト用シソーラスとして，自動的に構築することを試

† 豊橋技術科学大学 情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology

みる.

関連語リストを構築するためには, 二つの問題がある. 一つ目の問題は, テキストから未知語, 既知語を問わず関連語リストの対象となる単語の特定である. 対象となる単語の特定をするためにはまず, テキストの語分割を行わなければならない. しかし, 日本語には語の境界がないために, 語分割を正しく行うことが難しい. また, 語分割された単語をすべて関連語リストの対象とすることは実際的ではないため, 対象とすべき単語の特定が必要である. そこで, 辞書を用いずに文字列の頻度情報のみで語分割を行い, テキスト中のキーワードを抽出できるシステムを利用することにした[4),5)]. 本研究では, このシステムから得られるキーワードを関連語リストの対象とすることによって, テキスト中の重要語句の関連語を抽出することを試みる.

二つ目の問題は, 抽出するシソーラスに登録するべき単語の定義である. シソーラスには, 同義語, 類義語, 上位語, 下位語などさまざまなものがある. 本研究では, テキスト集合中で同じように使用される単語をシソーラスに登録することにした. シソーラスでもっとも重要な単語は同じ意味で用いられる単語や似た意味で用いられる単語である. たとえば, 二つの単語の関係が同義関係である場合, 一方の単語を含む文に対して, その単語の部分を他方の単語に置き換えても, その文の意味は同じになる. 言い換えると, 同じように使用される二つの単語は同義語であり得る. 同義語以外の上位語, 下位語についても同じように使用される傾向があるのではないかと考えられ, そこで, 本研究では, 関連語をテキスト集合中で同じように使用される単語と定義実際に抽出された単語はどのようなものであるかを分析する.

2. シソーラス自動構築手法

本研究では, 以下に示す行程を経て, 未踏テキスト用シソーラスとなる関連語のリストを生成する. すべての行程で辞書は用いない.

(1) テキスト集合から単語を特定し, 関連語リストの対象となるキーワードの集合を求める.

(2) キーワード集合から作成されたキーワードの組について関連語の組であるかを判定し, 関連語リストを求める.

以下の節では, この二つの行程を順に説明する.

2.1 単語の特定

第一の行程は関連語リストの対象となる単語の特定である. 本研究では, 未踏テキストに含まれる未知語を理解することに役立つ関連語の発見を目的とするため, テキストにある未知語, 既知語を問わず単語を特定しなければならない. しかし, 日本語には語の境界がないため, 日本語テキストは計算機にとって処理しにくいという問題がある. このため, 関連語のリストの対象となる単語, 特に未知語の特定に失敗する場合が多い. そこで, 本研究では既存の未踏テキスト中のキーワード抽出システムを利用する[4),5)]. このシステムは辞書を用いず, テキストの部分文字列から概念を示す単語と判定できる文字列を頻度情報のみで特定し, その単語をキーワードとして抽出するシステムである. 本研究では, このシステムを用いて抽出したキーワードを関連語リストの対象となる単語として扱う.

本研究で用いるキーワード抽出システムは, 対象とするテキスト集合から文字列の頻度情報のみでキーワードを抽出するシステムである. 文字列の出現頻度は, 統計的に言語を処理するときの基本の統計量である. 情報検索においても, 「ある単語がドキュメントに現れる確率」に関する情報量で重みをつけると性能が向上することが知られている. 出現確率の $P(1\text{回出現})$ を推定するには以下の式が使われる.

$$P(1\text{回出現}) = \frac{DF(x)}{N} \tag{1}$$

また, adaptation[3)] として知られる統計量があり, 「ある単語が一つのドキュメントに現れたという条件で, 同じ単語がもう一度出現する $P(2\text{回出現}|1\text{回出現})$ の推定値である. この確率を推定するために, 対象の文字列 x に関して, 「文字列 x を含むドキュメントの数:$DF(x)$」と「その文字列 x を 2 回以上含むドキュメントの数:$DF_2(x)$」を数え上げる. そして, ベイズの規則を考慮した式により推定する. ここで, N はテキスト集合に含まれるテキストの総数とする. 文字列の出現がポアソン分布に従うとすれば, $P(2\text{回出現}|1\text{回出現})$ は $P(1\text{回出現})$ の値と等しくなるはずであるが, 英語のキーワードの $P(2\text{回出現}|1\text{回出現})$ は, $P(1\text{回出現})$ には依存せず一定の値となることが報告されている[3)].

定義 2.1 adaptation

$$\hat{P}(2\text{ 回出現}\,|1\text{ 回出現}) = \frac{\hat{P}(2\text{ 回出現} \wedge 1\text{ 回出現})}{\hat{P}(1\text{ 回出現})} = \frac{\hat{P}(2\text{ 回出現})}{\hat{P}(1\text{ 回出現})} = \frac{DF_2(x)/N}{DF(x)/N} = \frac{DF_2(x)}{DF(x)} \quad (2)$$

そこで, 文献 3 では語の境界がない言語におけるキーワードの $P(2$ 回出現 $|1$ 回出現$)$ と $P(1$ 回出現$)$ を分析したところ, 実際のコーパスでは $P(2$ 回出現 $|1$ 回出現$)$ は $P(1$ 回出現$)$ に比べ大きく, キーワードである文字列（たとえば,「自立移動ロボット」）の $P(2$ 回出現 $|1$ 回出現$)$ はそのキーワードの次の文字も含む文字列（「自立移動ロボットに」）の $P(2$ 回出現 $|1$ 回出現$)$ に比べ非常に大きいことを観測した. この観測を基に, キーワードらしい文字列の推定を行い, 辞書を用いずに未踏テキスト中のキーワード抽出を実現している. 詳細は文献 3,4 に譲る.

2.2 関連語の判定

第二の行程はテキスト集合から候補の組が関連語の組であるかどうかを判定する. 通常, 関連語を取り出す手法は単語の出現分布が類似しているかどうかを判定し関連語を取り出すが, 本研究では, 単語の前後に接続している文字列の関係を基に関連語の組であるかどうかを判定する方法を検討する.

本研究では, 未踏テキストの理解に役立つ関連語リストを生成する. 未踏テキストを理解するためには, テキスト中の未知語を理解する必要がある. 未知語を理解することにもっとも役立つ関連語は同じ意味で用いられる単語や似た意味で用いられる単語である. そこで, 関連を判定する方法として, 二つの単語が対象とするテキスト集合中で同じように使用されるかどうかを調べることにした. 具体的には, 二つの単語が対象とするテキスト集合中で前後に同じ文字列を持って出現するかどうかを調べることにした. たとえば,「年賀状を印刷しなければならない」という文がある場合,「印刷」を「プリント」に置き換えても同じ意味の文になる. この単語の置換えは「印刷」と「プリント」が同義語であるが故にできることである. このことから, 逆に, 二つの単語が前後に同じ文字列を持って出現するのであれば, 同義関係にあるのではないかと想定し, 関連語の判定を行うことにした. そこで, 本研究では, 関連語を前後に同じ文字列を持つ, テキスト中で同じように使用される単語と定義する. したがって, 本研究で得られる関連語はあくまで同じように使用される単語である. 判定に用いた式と関連語リスト $Pair$ は以下のように定義される.

定義 2.2 関連語の判定

x, y は文字列, a, b は判定される単語, $tf(z)$ はテキスト集合における文字列 z の総出現頻度, $df(z)$ は文字列 z が出現するテキスト数, N はテキストの総数とする. また, xay, xby はそれぞれ単語 a, b の前後に文字列 x, y を結合した文字列である.

$idf(z) = -\log(df(z)/N)$ のとき,

$$score(a,b) = tf(xay)idf(xay)tf(xby)idf(xby), \quad (3)$$

$$Pair = \{(a,b) \mid tf(xab) > 1 \wedge tf(xby) > 1 \wedge score(a,b) > \alpha\} \quad (4)$$

$score(a,b)$ には, 語の特徴度を表し, 語が特徴的に多く出現することの数量的な評価になっていると考えられる $tf \cdot idf$[2] を採用した. これは, $IDF(z) = 0$ ならばすべてのテキストに出現し, $TF(z) = 0$ ならばテキスト集合に一度も出現しないことを表し, 意味のある単語であれば両方を考慮したものであるという考えに基づいている. この $tf \cdot idf$ は, 現在の検索システムで広く用いられている指標であり, その有用性は経験的に実証されている. このため, 本論文では, 二つの単語の $tf \cdot idf$ を用いて, その値が, 閾値 α より高ければ, 二つの単語は関連語の組であると判定する.

3. 実　　験

実験では, 日本語で書かれた論文アブストラクト 33 万件[1] (以降, NTCIR1 と記述する) をテキスト集合として, 関連語リストの構築を行った. この節では, 行程ごとに得られた実験結果の一例を示す.

3.1 単語の特定

本システムではまず, テキスト集合から単語の特定を, 既存のキーワード抽出システム[4),5)] を用いて行う. その結果として抽出されたキーワードの集合を関連語リストの対象となる単語の集合とする. NTCIR1 はアブストラクトごとに識別番号, タイトル, アブストラクト, 著者によって

付与されたキーワードが含まれている．本論文の実験では，一つのアブストラクトが持つ内容をタブで区切り，一行にしたものをテキストとして使用する．図1にテキストの例を示す．

gakkai-0000000813　超音波伝播特性による超電導マグネットのクエンチ検出 (パルス点加熱による場合)　大型超電導マグネットでは、マグネットのクエンチの早期検出と常時診断が不可欠である。著者らはこのようなマグネットのクエンチ検出および診断に超音波技術の応用を試みている。今回は小型の超電導マグネットを試作し、超音波疑似白色雑音を用いてマグネットの内部状態の変化を伝播特性の変化として周波数領域で検討し、とらえる実験を行なった。その結果、マグネット内部に機械的・熱的変化がなければ、再現性のある特性が得られた。又、定常特性，過渡特性において、マグネット内の局部的温度上昇がある場合には、伝播特性の変化として、とらえることができた。　クエンチ検出 超電導マグネット 超音波

図1　テキストの例

Fig. 1　An Example of Text Data

図2にNTCIR1のテキスト集合から特定された単語の例を示す．先頭の数字は抽出元となるテキストの識別番号である．

813　超伝導 遮断器 電圧 遮断器 ガス遮断器 消弧エネルギー 超伝導 素子 超伝導 遮断器

814　超電導 巻線 リアクトル 限流器 電力系統 故障 限流器 超電導 巻線 リアクトル 限流器 装置 限流器送電線 零相 限流 二線 地絡 地絡 超電導 巻線 抵抗限流 限流器 超電導 巻線 リアクトル

図2　特定された単語の例

Fig. 2　Examples of Specified Word

次に，この単語集合から自動的に候補となる組の集合を作成する．本来なら，任意の単語の組を処理すべきであるが，実験では効率のため，関連があるテキストから単語を選び，関連語リストを構築した．図3に作成された候補の例を示す．

最後に，この候補集合に関して，テキスト中で同じように使用されるかどうかを調べることによって，関連語であるかを判定する．具体的には，

超伝導　遮断器
超伝導　ガス遮断器
超伝導　消弧
超伝導　エネルギー
超伝導　素子
超伝導　超電導
超伝導　巻線

図3　作成された候補の例

Fig. 3　Examples of Created Candidate

対象となる単語の前後に接続されている文字列が一致するかどうかを調べる．このとき，調べる前後の文字列の文字数によって，関連語リストは異なる．本論文では，それぞれ前後4文字について実験を行った．図4に，実際に抽出された関連語リストの一部を示す．各行の二番目にある実数は関連語の組が持つスコアである．

4. 分　析

この節では，前節に示した実験から得られた関連語リストに含まれる関連語の組について分析する．実験では，以下のような特徴を持つ関連語が抽出された．

- 同じ意味を表す単語
  - 表記の揺れを含む単語
  - 省略形の単語
  - 外来語のために表記が異なる単語
  - 文字種（漢字，ひらがな，カタカナ）が異なる単語
  - 文字コードが異なる単語
- 似た意味を表す単語
- 上位概念または下位概念を表す単語
- 同じ上位概念を持つ単語
- 反対の意味を持つ単語

図4に示す関連語の組は33番まではスコアの高い組から正しく切り出されていない単語の組を約30%削った上位リストで，34番以降は上位ではないが，特徴的なものである．

この図において，同じ意味を表す単語を持つ組は9,10,19,24,26,28,31, 35,37,39,40,41である．このうち，長音記号や区切り文字の有無による表記の揺れを含むものは10,24,31である．省略形によるものは19,28であるが，テキスト集合を見ると，7の組は「移動体衛星通信」の省略形として「衛星通信」が用いられている文章が多かった．外来語であるために表記が異なるものは24,37,39である．外来語は英単語のカタカナ読みを表記するため，人によって表記が異なる場合がある単語

である．たとえば，「シミュレーション」は，人によっては「シュミレーション」と表記される場合がある．この表記の違いは表記の揺れの一種と考えることもできる．文字種（漢字，カタカナ，ひらがな）が異なるものは 9,40 である．これは，様々な人によって記述されたテキスト集合を対象としているため，漢字で表記する者もいれば，カタカナで表記する者もいるということである．現在，検索システムにおいて，ユーザが入力するキーワードはたいてい一つの文字種で表記したものであり，ユーザがこの表記の違いを知っている場合，より検索性能を上げるために，ユーザ自身で別の文字種でキーワードを入力し，検索範囲を広げることが行われている．このことから，このような関連語を得ることができれば，検索に役立つと考える．文字コードが異なるものは，35,41 である．41 を示した理由は，35 の単語の前に接続される文字列が長くてもある程度のスコアを持つことを示すことによって，この関連語の組の出現が稀ではないことを示したかったためである．これは，人によって表記に用いる文字コードが異なる場合があるということを示している．また，文字コードの違いも表記の揺れの一種と考えることができる．

似た意味を表す単語を持つ組は 16,17 である．これらは，表記からもなんらかの関係があるように見えるが，実際にテキスト集合を見ると，ほぼ同じ意味を表す単語として使われていた．

上位下位概念の関係にある組は 1,6,7,18,20,23,27,29,30,32,34 である．　このうち，7,23,27,29,34 は，テキスト集合を見ると，同じ意味で用いられたり，似た意味で用いられている場合があった．このことから，辞書に記載されている一般的な関係を用いるのではなく，対象とするテキスト集合によって，一般的な関係とは異なるそのテキスト集合特有の関係を単語の組が持っている場合があるということがわかる．

同じ上位概念を持つ単語の組は 3,8,12,15,21,33,36,38 である．日々新しく生成される単語は，既知語に他の単語を結合したものがほとんどである．これは，新しい概念には必ず上位概念があり，その上位概念を基に新しい概念を表現する単語を生成するためだと考えられる．このことから，関連語の組が既知語と未知語である場合，既知語の上位語をたどることによって，未知語を理解することが可能となると考える．

反対の意味を持つ単語の組は 4,42 である．これは，同義語と同じく，一方の単語を他方の単語に置き換えても，文が持つ意味は変わってしまうが，文を作ることができる．このため，本論文で定義した関連語となる．これは，同義語や類義語ではないが，反対の意味を持つ既知語を知ることによって，対する未知語を理解することに役立つと考える．

上記の特徴で分類されない 2,11,13,14,22 はどれも実際に関連を持つ単語の組である．本研究では，テキスト集合中で同じように使用される単語の組を関連語と定義している．このため，定義式を満たす同じように使用される単語の組はすべて関連語の組としている．このような関連語の組が，情報検索に役立つかどうかの評価は今後の課題とする．

5. ま と め

未知語，既知語を問わず単語を未踏テキスト中のキーワードとして抽出し，かつその単語を理解することに役立つ関連語リストを未踏テキスト用シソーラスとして自動構築する手法を提案した．今後の課題はこの関連語リストが実際に情報検索に有効であるかどうかを評価することである．

参 考 文 献

1) Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuji Kato, and Souichiro Hidaka, Overview of IR Tasks at the First NTCIR Workshop, *Proceedings of NTCIR1 Workshop*, Vol.1, pp.11–44, 1999.
2) 相澤彰子，語と文書の共起に基づく特徴度の数量的表現について，情報処理学会論文誌，Vol.41, No.12, pp.3332–3342, 2000.
3) Kenneth W. Church, Empirical Estimates of Adaptation, Coling2000, pp.180–186, 2000.
4) 田中路子，武田善行，仲村大也，山本英子，梅村恭司，純統計処理によるキーワードの抽出実験，第 *42* 回プログラミング・シンポジウム報告集，pp.155–158, 2001.
5) 武田善行，梅村恭司，キーワード抽出を実現する文書頻度分析，計量国語学，第二十三巻二号，pp.65–90, 2001.

番号	$score(a,b)$	a	b
1	338092.2306	アレー	フェーズドアレー
2	277031.8131	エキスパート	ネットワーク管理
3	168482.4400	ＩＩ型	Ｉ型
4	133381.3200	カチオン	アニオン
5	118208.7033	制振	制震
6	111506.7882	ＢＩＣＭＯＳ	ＣＭＯＳ
7	72908.4996	移動体衛星通信	衛星通信
8	70628.1898	ＳＹＮＥＣＨＯＣＯＣＣＵＳ	ＳＹＮＥＣＨＯＣＹＳＴＩＳ
9	66035.1475	ファイバ増幅器	ファイバアンプ
10	52519.4738	サーバ	サーバー
11	52354.3399	マルチビーム	フェーズドアレー
12	51017.1954	添加光ファイバ	ドープ光ファイバ
13	50872.8078	オペレーション	ネットワーク管理
14	47142.8378	衛星放送	平面アンテナ
15	43216.6650	ＩＩＩ型	Ｉ型
16	42199.1381	アーム	マニピュレータ
17	41760.2269	せん断強度	せん断耐力
18	40607.2778	円形鋼管	鋼管
19	40324.7238	ＣＶＤ法	ＣＶＤ
20	39730.5653	イソプレン	ポリイソプレン
21	38383.6836	ケイ酸	ホウ酸
22	37785.8253	サブバンド	帯域分割
23	36773.1338	電導体	伝導体
24	32821.6334	フェーズドアレー	フェーズドアレイ
25	32300.1363	平面アンテナ	円偏波
26	30950.9167	柱はり	柱・はり
27	28612.9366	超電導	超伝導
28	27977.0477	メール	電子メール
29	26959.6749	アクチュエータ	モータ
30	26226.6518	温熱環境	熱環境
31	25458.9096	レーダー	レーダ
32	25316.0180	骨材	細骨材
33	20061.2575	角形鋼管	円形鋼管
34	17232.4043	ネットワーク管理	網管理
35	12013.1703	靱性	靭性
36	7500.9396	ＤＲＡＭ	ＳＲＡＭ
37	7240.2699	イソシアナート	イソシアネート
38	6511.7073	小学	中学
39	6299.5691	ウイルス	ウィルス
40	4000.8972	結合タンパク質	結合蛋白質
41	3661.8664	破壊靱性	破壊靭性
42	3348.7308	冬季	夏季

図4　関連語リストの一部

Fig. 4　A Part of Related Word List

本 PDF ファイルは 2002 年発行の「第 43 回プログラミング・シンポジウム報告集」をスキャンし、項目ごとに整理して、情報処理学会電子図書館「情報学広場」に掲載するものです。

この出版物は情報処理学会への著作権譲渡がなされていませんが、情報処理学会公式 Web サイトに、下記「過去のプログラミング・シンポジウム報告集の利用許諾について」を掲載し、権利者の捜索をおこないました。そのうえで同意をいただいたもの、お申し出のなかったものを掲載しています。

https://www.ipsj.or.jp/topics/Past_reports.html

過去のプログラミング・シンポジウム報告集の利用許諾について

情報処理学会発行の出版物著作権は平成 12 年から情報処理学会著作権規程に従い、学会に帰属することになっています。

プログラミング・シンポジウムの報告集は、情報処理学会と設立の事情が異なるため、この改訂がシンポジウム内部で徹底しておらず、情報処理学会の他の出版物が情報学広場 (=情報処理学会電子図書館) で公開されているにも拘らず、古い報告集には公開されていないものが少からずありました。

プログラミング・シンポジウムは昭和 59 年に情報処理学会の一部門になりましたが、それ以前の報告集も含め、この度学会の他の出版物と同様の扱いにしたいと考えます。過去のすべての報告集の論文について、著作権者（論文を執筆された故人の相続人）を探し出して利用許諾に関する同意を頂くことは困難ですので、一定期間の権利者捜索の努力をしたうえで、著作権者が見つからない場合も論文を情報学広場に掲載させていただきたいと思います。その後、著作権者が発見され、情報学広場への掲載の継続に同意が得られなかった場合には、当該論文については、掲載を停止致します。

この措置にご意見のある方は、プログラミング・シンポジウムの辻尚史運営委員長 (tsuji@math.s.chiba-u.ac.jp) までお申し出ください。

加えて、著作権者について情報をお持ちの方は事務局まで情報をお寄せくださいますようお願い申し上げます。

期間： 2020 年 12 月 18 日 ～ 2021 年 3 月 19 日
掲載日： 2020 年 12 月 18 日

プログラミング・シンポジウム委員会

情報処理学会著作権規程
https://www.ipsj.or.jp/copyright/ronbun/copyright.html