自然言語処理 105－3
（1995．1．20）

連語検索結果の評価手法と提示優先順位決定アルゴリズム

青山 昇一* 徳田 克己* 塩見 隆一* 辻 敦生** 藤田 公一* 杉村 領一*
*松下電器産業（株） 情報通信研究所 **松下電器産業（株） 情報機器事業部

英文を読むためのツールとして単語辞書検索と同時に、イディオム・構文知識などの連語辞書データを提示する連語辞書検索システムを実現した。しかし、このシステムは非常に多くの連語を検索し、利用者が英文を読むのに必要な連語を見つけることが難しいことがわかった。そこで、この問題を解決するため（1）検索された連語に優先度を計算し、優先順に連語をを表示する手法（2）検索する連語数を削減する手法を考案した。連語に優先度をつけることで、必要な連語を上位6位以内に収めることができた。また、必要な連語を失うことなく検索する連語数を５０％削減することができた。

An Algorithm for Deciding the Order of Presentation for Results of Phrase Dictionary Retrieval

Shoichi Aoyama* Katsumi Tokuda* Takakazu Shiomi*
Atsuo Tsuji** Kimikazu Fujita* Ryoichi Sugimura*

*Information and Communications Technology Laboratory,
Matsushita Electric Industrial Co., Ltd.
**Information Equipment Division, Matsushita Electric Industrial Co., Ltd.

We developed a Phrase Dictionary Retrieval System which presents idioms and syntactical knowlege as a tool to read English. However, the system retrieves many phrases and an operator cannot find the phrases which he required. In order to solve this problem, we developed: (1) An algorithm for deciding the order of presentation for results of phrase dictionary retrieval, (2) A method of decreasing the number of the phrases retrieved from the dictionary. Using the first method, all the required phrases are in the top six. And using the second method, we decreased the number of phrases by 50% without losing the required phrases.

1 はじめに

近年、英和辞典が電子手帳をはじめとする携帯情報機器に搭載され、英文を読む際に手軽に利用できるようになった。しかし、単語の訳を組み合わせただけでは、英文の意味を読みとることは困難である。この事実は、大多数の利用者にとって電子辞書が英文読解のツールとして不十分であることを意味する。一方、英語を日本語に翻訳する機械翻訳システムがある。しかし、現状の機械翻訳システムを実用的に使用するためには、ユーザー辞書への辞書登録やユーザーに応じたシステムのチューニングなどが必要であり、個人ユーザーが手軽に利用することはできない。電子辞書と機械翻訳システムの中間に位置するシステムが必要であると考えられる。

そこで我々は、連語辞書検索システムを提案する。連語辞書検索システムは、単語辞書検索と同時に、従来の電子辞書が提示できなかったイディオム・構文知識などを提示し、利用者の英文読解を容易にするものである。本稿では、

1. 連語辞書データの検索手法
2. 検索した連語辞書データの提示優先順位決定アルゴリズム
3. 連語辞書検索システムの評価

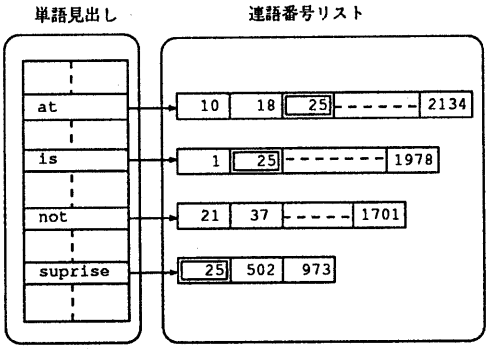
について報告する。

2 連語辞書検索システム

連語辞書検索システムは、2単語以上で構成される連語見出しとその訳情報で構成された連語辞書データ（以下、連語と記述する）を検索するものである。基本的に2～3単語で構成される複合語やイディオムを検索することを想定しているが、諺や文などを取り扱うこともできる。連語辞書の構成及び検索アルゴリズムは、例文検索システム [1, 2] や全文検索システム [3] で用いられているものと同様である。

2.1 連語辞書構成

図1は、辞書とインデックスの構成図である。連語は、「連語番号」「連語見出し」「連語訳語」で構成され、連語辞書に格納される。単語見出しは、連語見出し中に含まれる全単語を集めたものである。連語番号リストは、各単語見出しに対応した連語を検索するためのインデックスであり、対応する単語を連語見出しに含んでいる連語の連語番号を集めたものである。



連語辞書

連語番号	連語見出し	連語訳語
1	be able to	～できる
25	be suprised at	～に驚く
26	be amazed at	～にびっくりする
27	at once	直ちに、同時に

図 1: 連語辞書構成図

連語番号リストと連語辞書の関係を具体的に説明する。図1中では、連語辞書に

連語

連語番号	連語見出し	連語訳語
25	be suprised at	～に驚く

が格納されている。連語見出しの各単語の原型単語 “is, suprise, at” は、単語見出しに格納されている。この各単語に対応した連語リストには連語番号25が格納されている。

2.2 連語辞書検索アルゴリズム

入力された文に対して、以下の手順で連語辞書検索を行なう。

1. 検索した連語を格納する辞書バッファを用意する。
2. 入力文の形態素解析を行ない単語を抽出する。
3. 入力文中から2単語を検索単語として選択する。
4. 検索単語の連語リストから共通する連語番号の連語を連語辞書から検索し辞書バッファに格納する。
5. 検索単語と連語見出し中の単語で一致したものに印をつける。
6. 4,5を、入力文中の全2単語の組合せに対して行なう。なお、辞書バッファが同一辞書データを格納している場合は、検索単語と連語見出し中の単語の一致を調べ、印の更新のみ行なう。

2.3 連語辞書検索例

具体例として、図1の連語辞書に対して

入力文

I was not surprised at him.

が入力された場合の連語検索を説明する。最初に形態素解析が行なわれ、

6つの単語

(1)I (2)be (3)not
(4)surprise (5)at (6)him

が得られる。この6単語から全2単語の組合せ15通りに対して、共通する連語番号を連語リストから抽出する。

表1は、2単語の単語組から抽出された連語番号を表している。7通り目で初めて連語番号25が抽出され、辞書バッファには25番の連語が格納され、一致した単語”be” と”surprised” に印が付けられる。図2 （a）は、この時の辞書バッファの状態を表している。

No	単語組	連語番号	No	単語組	連語番号
1	(1) (2)	なし	9	(2) (6)	なし
2	(1) (3)	なし	10	(3) (4)	なし
3	(1) (4)	なし	11	(3) (5)	なし
4	(1) (5)	なし	12	(3) (6)	なし
5	(1) (6)	なし	13	(4) (5)	25
6	(2) (3)	なし	14	(4) (6)	なし
7	(2) (4)	25	15	(5) (6)	なし
8	(2) (5)	25,26			

表 1: 単語組と抽出された連語番号

8通り目では連語番号25と26が抽出される。25番の連語は既に辞書バッファに存在するため、一致した単語"at" に印がつけられる。また、26番の連語は新たに辞書バッファに格納され、一致した単語に印が付けられる。図 2 (b) は、この時の辞書バッファの状態を表している。

最後に13通り目で、連語番号25が再度抽出される。この連語は既に辞書バッファに格納されており、連語見出しの全単語には既に印が付けられているため、辞書バッファは図2 (b) の状態から変化しない。

(a) 7通り目の終了

be surprised at　～に驚く

(b) 8通り目の終了

be surprised at　～に驚く
be amazed at　～にびっくりする

図 2: 連語辞書検索語の辞書バッファ (連語見出しの下線は検索単語の一致を表す

2.4 連語辞書データ

連語辞書データは、デイリーコンサイス英和辞典[4] から抽出した。2単語以上で構成される辞書見出しから約8000、本文中から約12000を抽出し、全連語は約20000である。抽出した連語の連語見出しに含まれる平均単語数は 2.4 語であった。なお、本文中から抽出された連語辞書データには、諺や例文が含まれている。

3 連語辞書検索システムの問題点

連語辞書検索システムを試作し、実際に文を入力して連語の検索を行なった。入力文には、日本電子工業振興協会で作成された英日機械翻訳システムの評価を行なうための評価文[5] を用いた。表2は、評価文と検索結果をまとめたものである。評価文309文に対して検索された総連語は8119であり、1評価文あたり約26の連語が検索された。また、1評価文に対して検索された連語の最大数は363であった。

検索された連語の中で、**有用連語** (評価文の意味に合致した連語) と**参考連語** (評価文を読むために参考となる連語) の個数について調べた。表 2 における有用連語・参考連語の判定は人間が行なったもので、有用連語数は70、参考連語数は28であった。図3は、人間によって判定された有用連語と参考連語の例である。この結果から、本試作システムでは有効連語・参考連語数と比較して検索される連語数が多過ぎることが判明した。

評価文数	309
評価文1文あたりの平均単語数	6.7
検索された連語総数	8119
評価文1文あたりの平均連語数	26.3
評価文1文あたりの最大連語数	363
有用連語数	70
参考連語数	28

表 2: 連語辞書検索結果

評価例文	I spend a lot of money on books. 私は本に多くのお金を費やす
有用連語	a lot of <話>たくさんの
評価例文	They told me how much a car cost. 彼らは私に車がいくらするかを教えた。
参考連語	How much is it ? (値段は) いくらですか?

図 3: 評価例文に対する有用連語と参考連語の例

図4は、連語を検索した順に提示した場合の、提示順位に対する有用連語と参考連語の分布について

調べたものである。当然のことではあるが、有用連語や参考連語が上位に提示されるとは限らない。利用者が有用連語や参考連語を得るには、検索された全連語を見る必要がある。この結果から、利用者が本試作システムから有用連語や参考連語を容易にみつけることが難しいことが判明した。

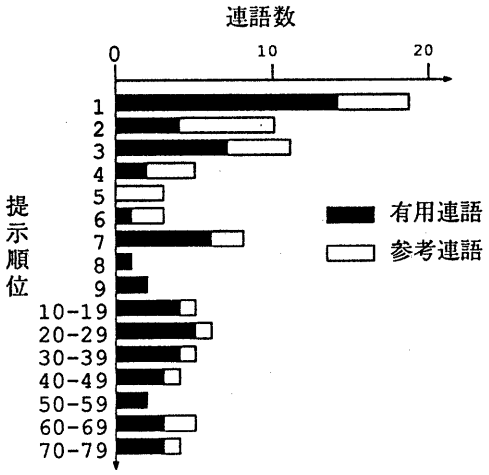


図 4: 検索順位～有用・参考連語数

4 提示優先順位決定アルゴリズム

連語辞書検索システムの問題点を解決するために、以下の処理を行なった。

1. 検索された連語の優先度を計算し、優先順に連語を表示することによって、有用連語や参考連語を上位に提示する。
2. 検索する連語数を削減する。

我々は提示する連語の優先順位決定に関して、5つの指標を組み合わせる手法と、単語の出現確率を用いる手法を考案した。また、検索連語数の削減に関しては、単語の出現確率を用いる手法を考案した。

4.1 5指標を用いた優先順位決定法

検索された全連語に対して5つの指標の値を計算し、その値の組合せで検索された連語を並び換える手法である。5つの指標は以下の通りである。なお、図5に、各指標の算出例を示しておく。

1. **連語一致数**
   入力文中の単語と連語見出し中の単語で一致している単語の総数。単語数が多いほど優先度は高い。
2. **連語一致率**
   連語一致数を連語見出し中の総単語数で割った値。一致率が高いほど優先度は高い。
3. **入力文中不要語数**
   連語見出し中の単語と一致する入力文中の単語の間にある不要な単語の総数。不要語数が多いほど、優先度は低い。
4. **連語中不要語数**
   入力文中の単語と一致している連語見出し中の単語の間にある不要な単語の総数。不要語数が多いほど、優先度は低い。
5. **変化形単語数**
   入力文中の単語の中で原型抽出を行なったことで、連語見出し中の単語と一致した単語の総数。単語数が多いほど優先度は低い。

入力文	I was not surprised at him.
連語1	be surprised at
連語2	be amazed at

指標	連語1	連語2
連語一致数	3	2
連語一致率	100%	67%
入力文中不要語数	1	2
連語中不要語数	0	1
変形単語個数	1	1

図 5: 評価指標の算出例

4.2 単語の出現確率を用いた優先順位決定法

我々は、連語見出しを構成する単語には、その連語のキーとなる重要な単語が存在すると考えた。本手法では、検索された連語に対する優先度 U を

$$U = \frac{\sum_{j \in C} W_j}{\sum_{k=1}^{N} W_k}$$

と定義した。ここで、W_i は連語見出し中の単語 i の重要度、N は連語見出し中の総単語数、C は入力文中の単語と一致する連語見出し中の単語の集合である。

単語 i の重要度 W_i は、

$$W_i = \frac{1}{P_i}$$

と定義した。ここで、P_i は単語 i の出現確率である。これは、単語 i の情報量 $I_i = \log \frac{1}{p_i}$ の対数を外

したものである。情報量を用いてもよいが、計算の簡単化のためこの式を採用した。

優先度Uは、入力文中の単語と一致する連語見出しの単語の重要度の総和を、連語見出しの全単語の重要度の総和で割った値である。重要度が高い連語見出しの単語が入力文中の単語と一致すると、優先度Uの値が高くなる。

図6は、優先度Uの算出例である。

単語	be	surprize	amaze	at
重要度	100	9600	24000	134

入力文	I was not surprised at him.
連語1	be surprised at
連語2	be amazed at

連語1　$U_1 = \frac{100+9600+134}{100+9600+134} = 1.00$

連語2　$U_2 = \frac{100+134}{100+24000+134} = 0.01$

図6: 単語の重要度を用いた優先度の算出例

4.3 単語の出現確率を用いた検索連語削減法

連語見出し中の単語の重要度に対して、閾値を導入する。連語辞書を検索するための入力文中の2単語の組合せの中で、2つの単語の重要度が設定された閾値より小さい場合は、その2単語の組合せでは連語辞書検索を行なわない。これは、単語の情報量を使用して、設定された情報量より小さな情報量の単語組で検索を行なわないことと等価である。

この手法を用いると、重要度の低いbeとatなどの単語組による検索を除くことができるため、2.2節の具体例で検索されていたbe amazed atのような不必要な連語が検索されなくなる。

5 評価実験

以下の評価実験では、連語検索を行なう入力文として、日本電子工業振興協会で作成された英日機械翻訳システムの評価文を用いた。

5.1 優先順位決定法

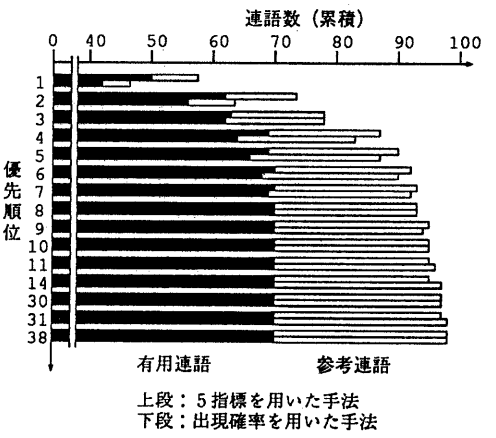
連語検索を行なった後、2つの手法で連語を優先順に並べ換え、有用連語と参考連語の順位の分布を調べた。

5指標を用いた手法では、

1. 連語一致率
2. 連語一致数
3. 入力文中不要語数
4. 連語中不要語数
5. 変化形単語数

の順に適用するものとし、ある指標で優先順が等しい場合には、優先順が決まるまで次の指標を順に用いて比較を行なう方法を用いた。

図7は、評価文309文に対する有用連語と参考連語の順位別の分布を表している。5指標を用いた優先順位決定法は1位と2位の個数が多い。単語の出現確率を用いた優先順位決定法は14位以下の累積連語数が多い。また、有用連語と参考連語の低順位に注目すると、出現確率を用いた優先順位決定法が31位で、5指標を用いた優先順位決定法の38位を上回っている。



上段：5指標を用いた手法
下段：出現確率を用いた手法

図7: 優先順位～有用・参考連語数

5.2 検索連語削減法

単語の重要度に対して閾値を変化させ、その際に検索される連語数とその中に存在する有用連語数及び参考連語数を調べた。また、設定された閾値よりも重要度の低い単語も調べた。表3は実験結果をまとめたものである。

閾値を大きくすることにより、閾値より重要度の低い単語数が増加し、検索される連語数が減少する。有用連語及び参考連語は、閾値が$\frac{1}{300}$より低い場合は減少しない。

5.3 優先順位決定法と検索連語削減法の融合

検索連語削減法の適用後に優先順位決定法を用いて連語を優先度順に提示するシステムを作成し、両手法を同時に用いた場合の効果を調べた。

表3の実験結果を参考にし、検索連語削減法の閾値には有用連語及び参考連語の検索洩れがない$\frac{1}{300}$

閾値	検索連語数	有用参考連語数	閾値より低い単語
なし	8119 (100%)	98(70)	
$\frac{1}{400}$	4334 (53%)	98(70)	of, in, one's, to, on, out, up, be, and
$\frac{1}{300}$	3858 (48%)	98(70)	上記に加えて at, with, for, off
$\frac{1}{250}$	3676 (45%)	97(69)	上記に加えて as, make, take
$\frac{1}{200}$	2431 (30%)	89(62)	上記に加えて go, by, have it, oneself

表 3: 閾値と連語数（括弧内は有用連語数）

を採用した。優先順位決定法としては、単語の出現確率を用いた優先順位決定法を採用いた。

図 8 は、検索連語削除法と優先順位決定法を融合した場合と、優先順位決定法のみの場合の有用連語及び参考連語の順位別分布を表したものである。2 つの手法を融合することにより 1 位の有用連語・参考連語数が増加する。

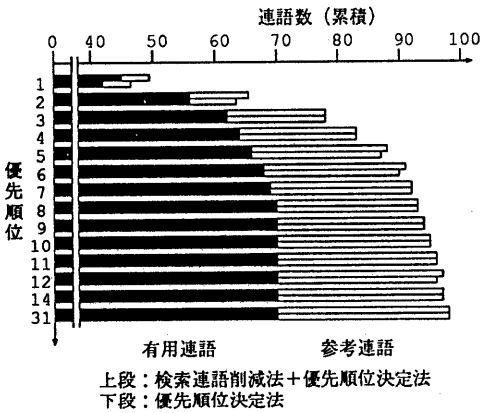


図 8: 優先順位～有用・参考連語数

6 考察

優先順位決定法を用いて連語を提示することにより、有用連語や参考連語を上位候補として提示することが可能となった。とりわけ有用連語は、5 指標を用いた優先順位決定法では上位 6 位以内に含まれ、また、単語の出現確率を用いた優先順位決定法でも上位 8 位以内に含まれることがわかった。本実験では、5 指標を用いた優先順位決定法が、若干良好な結果を出している。しかしながら、両手法の優劣を決定するほどの差はなく、今後、より多くの評価文に対して評価実験を行なう必要があると思われる。

検索連語削減法における閾値の範囲は、有用連語や参考連語の検索洩れがない状態までとするのが妥当であると考える。表 3 の結果から、閾値は $\frac{1}{300}$ が妥当であり、削減できる検索連語数は約５０％程度である。この程度の連語削減だけでは、利用者が有用連語や参考連語を容易に利用できるとは限らず、検索連語削減法は問題点を解決する有効な手法とは言い難い。しかし、検索連語削減法は、検索連語数を５０％削減すると同時に、検索時間や検索用辞書バッファ領域の削減を行ない、コンパクトな連語辞書検索システムの構築に寄与する。また、表 8 の結果より、優先順位決定法と同時にシステムに組み込むことで優先順位決定法の効果を向上させることが可能である。

7 まとめ

以上、本稿では英文読解のための連語辞書検索システムを提案した。今後は、このシステムを利用して、英文読解を行なった場合の効果について評価を行なう予定である。

謝辞

デイリーコンサイス英和辞典第５版の電子化データの使用を許可して下さった株式会社三省堂に感謝致します。

参考文献

[1] 隅田英一郎, 堤豊:" 翻訳支援のための類似用例の実用的検索法," 電子情報通信学会論文誌,Vol.J74-D-II,No.10,pp.1437-1447,1991.

[2] 中村直人:" 用例検索翻訳支援システム," 情報処理学会全国大会講演論文集, 4E-5,pp.357-358, 1988.

[3] 菊池忠一:" 日本語文書用高速全文検索の一手法," 電子情報通信学会論文誌, Vol.J75-D-I, No.9, pp.908-916, 1994.

[4] 三省堂編修所:" デイリーコンサイス英和辞典第５版," 三省堂, 1990.

[5] 日本電子工業振興協会:" 自然言語処理の動向に関する調査報告書," 94-計-4, 1994.