自然言語処理 121－10
情報学基礎 47－2
（1997．9．11）

長い日本語表現の高速類似検索手法

田中英輝
NHK 放送技術研究所
（現在　ATR 音声翻訳通信研究所）
tanakah@itl.atr.co.jp

内容梗概　著者らは用例提示型日英翻訳支援システムを開発している．この中にはユーザが入力する日本語表現の類似表現を検索し，これを含む日本語文と英訳を提示する機能がある．著者らの日本語データベースの文は平均長 88.9 文字と長い．このような長文を対象に日本語表現の類似検索を行う場合，従来のキーワードを使った Boolean 検索は適切でない．なぜならデータベースの一文中に同一キーワードがいくつも出現するため雑音を検索しやすいからである．特に入力が長いとこちらにも同一キーワードが出現して問題となる．これに対し著者らは入力キーワードの語順とその間隔を考慮した検索手法を提案する．これは構文解析を行わず近似的に構文を考慮する手法である．本稿では (1) 提案手法，(2) Boolean 検索，(3) キーワードの語順を考慮する手法を考察して実験的に比較する．そして提案手法の検索結果の適合性が最も高いことを示す．さらに本手法が結果の提示手法としても優れていることを示す．

An efficient Way of Gauging Similarity between Long Japanese Expressions

Hideki Tanaka
NHK Science and Technical Research Labs. (presently with ATR)
tanakah@itl.atr.co.jp

Abstract　We are developing a J-E news article browser for translators. The system accepts a Japanese expression as an input, and then targeting all past Japanese articles it searches for expressions similar to the one input. Finally English sentences corresponding to those Japanese sentences that include found expressions are displayed simultaneously with the Japanese.

The Boolean retrieval with keywords has been used for this kind of similar expression search since it runs fast. However this approach is not appropriate for our task. The Japanese sentence in our database is long whose average length reaches 88.9 Japanese characters. Then Boolean method captures many spurious sentences since such a long sentence quite often includes a same keyword several times.

We propose a retrieval method which takes the order of keywords and their positions into account. This method can approximate the syntactical similarity between expressions without parsing them. We compare the proposed method with the Boolean method and the Boolean method solely with keyword order. We report that our method showed the best precision among the three. We also point out our method's superiority as a way to present the retrieval result.

1 はじめに

著者らは用例提示型日英翻訳支援システムを開発している (熊野 97)．これはユーザの入力する日本語を日英対訳データベースで検索してその翻訳例を提示するシステムである．本稿では便宜的に入力の日本語を「表現」と呼び，検索対象の日本語文を「用例」と呼ぶ．すなわち「表現」を入力して「用例」を検索する．

著者らの用例，すなわち日本語ニュースの文は平均で 88.9 文字と長い (熊野 96)．このため従来のキーワードを使った Boolean 検索を採用すると特に長い入力表現で 2 章で述べる不都合が生ずる．本報告の主要な内容はこの解決法の提案である．

著者らはニュースの翻訳者を支援することを目的としている．そこで NHK の日本語ニュースとその英訳記事を使った対訳データベースをこの翻訳支援システムのために作成している．

この中で著者らが検索機能を設計する際に目標としたのは次の点である．

(1) 任意長の文字列の入力
一文字から文までの間の任意長の日本語文字列を入力表現として受け付ける

(2) 完全一致検索と類似検索の実現
短い表現，長い表現に対して有効な結果を得るため表記二手法を選択できるようにする

(3) 高速な検索
用例数は 100 万件程度を想定してこれを高速に検索する．検索精度が悪くても高速であればユーザは何度でも検索でき，所望の結果を得ることができる

(4) 理解容易な結果提示
結果をすばやく評価できるようにする．このためには直感的にわかりやすい基準で類似性を評価する

これらを実現するために著者らは次のような手法を採用した．

(5) 全文検索技術の採用
日本語のデータベースに対して，ポインタ表現による部分文字列インデックス (Nagao 94) を作成する．これにより任意長文字列の高速な完全一致検索はそのまま実現できる

(6) キーワードの共有数による類似性の評価
入力表現と用例の類似性の評価には構文解析を使う手法，シソーラスを使って意味的な類似性を考える手法がある．しかし著者らは長い用例に対して頑健な表層的手法を採用する．すなわち入力表現を形態素解析して，この中のキーワードを共有する数で入力と用例の類似性を評価する．キーワードの選択については後述する．照合したキーワードを結果に明示することでユーザが類似性を直感的に判定できる利点が生ずる

2 長文用例の類似検索の問題点

これまで提案されている代表的な類似用例検索手法はキーワードを使った Boolean 検索 (Salton 83) である．先に述べたように著者らの対象とする日本語の文は長い．このため Boolean 検索では下記の問題を生ずる．例えば「政府の作業」の類似用例を検索する場合，(「政府」AND「作業」) で検索すると下記をすべて検索する[1]．

例 1
外務省の橋本外務報道官も，きのうの記者会見で，「保証人委員会は一生懸命作業をしているが，ペルー政府と武装グループが，保証人委員会の努力を受け入れる所まで事態は進んでいない」と述べました．

例 2
この問題に関する自民党の対外経済協力特別委員会が今日午後開かれ，政府側は，「中国は去年七月に核実験を行なった後，今後の核実験を凍結すると表明しており，無償資金協力の再開に向けた準備作業を進めていきたい．」と述べました．

例 3
また池田外務大臣は，「日本政府とペルー政府との間は信頼関係が出来ている」と述べ，両国政府の間で緊密に連絡を取っていることを明らかにするとともに，今後の日本の役割について「関係国の間で，バラバラの対応にならないよう，国際社会が一致してペルー政府の進め方を支えていくことが重要だ．日本政府は，事件の解決に向けたペルー政府の作業がうまく運ぶよう，条件を整える努力をしてきており，今後はこうした努力が一層大切になる」と述べました．

尚，ユーザに類似性の根拠を示すため照合キーワードを強調して表示している．例 1 はキーワー

[1] 実際には完全一致で検索するべきである

ドの順序が逆転しており正解ではない．例2では語順は正しいものの係り受けが違っている．正解は例3に含まれている．しかし「政府」が6個所に出現しているため該当個所を見いだすのは容易ではない．

ここで使った表現「政府の作業」は短く同じキーワードはない．しかし長い入力表現では入力にも同じキーワードが何度も出現する．この場合，照合部分を把握するのはさらに困難になる．

結局Boolean検索は高速ではあるが，例1と例2で示したような雑音を拾いやすく，また正解であっても判定しにくい問題がある．

これを解決するには構文解析を利用する手法があるが現時点では精度と速度の点で採用しにくい．著者らはこれらの問題を構文解析せずに4章で示す近似的な手法で解決する．

3　システムの概要

3．1　用例データベース(日本語)

1996年3月から1997年2月までのNHKの日本語の記事を利用して作成した．

記事の数:　94,830件
文数(付加情報を含む):　1,615,119件
バイト数:　104 MB

この記事データベースに対してポインタ表現の部分列インデックス(Nagao 94)を作成して任意の文字列を高速に検索可能とした．検索結果は文字列が出現した記事，文内での位置である．

3．2　入力表現解析

入力表現を形態素解析して事前に指定した品詞の単語をキーワードとして抽出する．現在は自立語をキーワードに採用している．また活用自立語は可能な変化形に展開する．

4　類似検索

4．1　検索の流れ

全体的な類似検索手続を示す．ここで記事データベースの用例(文)の全体集合をSとし，また入力表現にN個のキーワードがあるとする．

```
while ( N > 0) {
    S からキーワードを N 個含む
      文集合 S(N) を検索; S(N) の提示;
    if ( ユーザが終了を指示 ) { 終了; }
    if ( S(N) = ø || ユーザが緩和を指示 ) {
        S <- S - S(N); N <- N - 1; }
}
```

すなわち最大i個のキーワードを含む用例の集合をi＝Nからi＝1までユーザの検索条件の緩和指示に従って行う．

尚，検索結果の提示を行う場合には用例中の出現キーワードを強調表示する．また用例が出現した記事もあわせて表示する．

4．2　キーワードの語順と変位の考慮

検索手続の中で，「Sからキーワードを N個含む文集合 $S^{(N)}$ を検索」する部分には「キーワードの語順と変位」を考慮した手法を採用している．これを手法(1)と呼ぶ．以下この手法の目的と効果を説明する．

次の入力表現を考えよう．

入力表現　　**A*B**A*C*

ここでA，B，Cはキーワード，*はそれ以外の単語とする．また簡単のため単語はすべて一文字とする．

語順を無視した手法，すなわち先に述べたキーワードを使ったBoolean検索(手法(2)と呼ぶ)では次の用例をいずれも第一候補として検索する．これらに順序はない．照合したキーワードは強調表示(下線)している．用例3と用例4にはAが3つあるが，この手法ではどの2つを照合したかを決めることができずあいまい性が残る．

用例1　　*A**B*C**A**
用例2　　*A A*B***C
用例3　　*A**A**B*A**C
用例4　　*A**A*B**A*C

これに対して語順を考慮した手法(手法(3)と呼ぶ)ではキーワードの語順の一致する上記の用例3と用例4だけを第一候補とする．

用例3　　*A**A**B*A**C
用例4　　*A**A*B**A*C

これらの用例にも順序はない．またこの手法でも先頭の2つのAのどちらが照合キーワードなのか決めることはできずあいまい性が残る．

提案手法(1)では変位を使って用例間に順序を付け，また照合キーワードのあいまい性を解消する．変位とは用例と入力のキーワード間隔のずれである．具体的には「用例中のキーワード対の間隔と入力のキーワード対の間隔の差の絶対値」である．例えば用例1のキーワード対(A, B)の間隔は5-2＝3である．入力中のキーワード対(A, B)の間隔は5-3＝2である．すなわち両者の変位は

|3-2|=1 となる．

照合キーワードのあいまい性は変位合計が最小となるキーワード対の組み合わせを求めることで解消する．また用例は変位合計の小さな順に提示する．現在の例では (A, B), (B, A), (A, C) の変位合計が最小になるキーワードを選択することになり，この値の小さな順に提示することになる．この結果，強調するキーワードと用例の提示順序は下記のように確定する．

用例 4　　*A**A*B**A*C　　(変位合計 0)
用例 3　　*A**A**B*A**C　　(変位合計 3)

この手法はキーワード対を一種の係り受けと捉えている．そして間隔が近い二つの係り受けは構文的に近いと仮定している．すなわちキーワード照合のあいまい性は，入力と用例の係り受け全体が最も近くなるキーワード対を選択することで解消していることになる．

手法 (1) と (3) が同じ入力キーワード群で検索する用例の数は上記のように常に一致する．手法 (1) はキーワード照合にあいまい性がない点と用例に順序がつく点が手法 (3) と異なる．

一般的に上記の例のように Boolean 検索 (手法 (2)) は同じ入力キーワード群に対して語順を考慮した手法 (1, 3) より多くの文を検索する傾向を持つ．ただし，キーワード数を 1 まで緩和して検索できる文の総数はいずれの手法も一致する．この意味でここで述べた 3 手法の再現率は等しい．

4．3　実装法

キーワード数の緩和を考慮した上で語順と変位を使って用例を効率よく検索するため，著者らは動的計画法を利用した手法を使っている．ここではその概要を示す．

(i) ノード集合の作成

図 1 の上段は入力表現のキーワード列とその出現位置をノードとしたノード集合である．下段には入力キーワードが用例 4 で出現した位置をノードとして表示しそれぞれにノード番号を付与している．用例 4 と同様のノード集合をキーワードを一つ以上含む用例で作成する (S に相当する)．

このノード間を，右方向に出現位置が増加するようにつないだ任意の経路が一つの照合キーワード群になる．

(ii) 検索

S のすべての用例を対象にキーワードが 4 つ

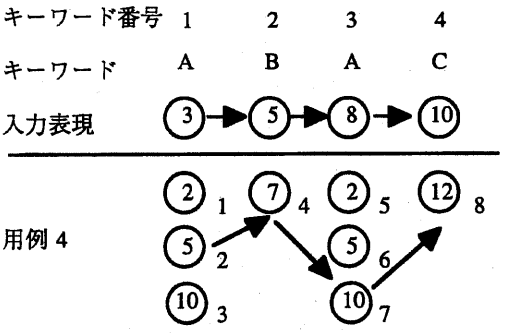

図1 検索の概要

，すなわち N = 4 で検索する．4 つのキーワードを含む可能性があるのはキーワード番号 1 の A を含む場合しかない．そこで用例 4 であればキーワード番号 1 の 3 つのノード (ノード番号 1, 2, 3) を開始点とした経路を動的計画法で評価する．この結果 (2->4->7->8) で変位合計 0 という最適解を得る．以上の処理をその他の用例についても実施する．そしてキーワードを 4 つ含んだ用例を変位合計で整列表示してそのノード集合を削除する (S <- S- $S^{(N)}$ に相当する)．またキーワード数が 4 に満たない解を得た場合はこれを記憶しておく．

(iii) 緩和

次に一つ緩和して キーワードを3 つ含む解を求める処理を行なう．キーワード数 4 の時に得たキーワード数 3 の解と，キーワード番号 2 のノード集合を開始点として得られる解がすべての解である．これらの解を (ii) と同様に求めて表示する．

以上のように本手法はキーワードを最大限含む解から順に解を求めている．多くの場合ユーザは緩和の途中で検索を打ち切るので，この順序で解を求めている．変位を考慮せず語順だけを考慮する手法も同様に実現でき，この場合はさらに高速である．

5　検索実験

5．1　検索時間

語順と変位を使った手法 (1)，語順だけを考慮した手法 (3)，語順を使わない Boolean 検索手法 (2) を対象に検索時間を評価した[1]．入力したのは 1997 年 3 月のニュース記事 500 件の各先頭行 500 行である．平均文字数は 92.7 である．評価は以下のようにして行った．

[1] 主記憶 1GB, SPECint92 = 202.9, SPECfp92 = 259.5

表1 検索時間の比較(秒)

手法	(1)	(2)	(3)
総時間	33,426.9	25,573.1	33,201.6
緩和平均	3.04	2.33	3.02

入力の各文を形態素解析して自立語キーワードを抽出する．そして各文で検索を行い，キーワード数1になるまで条件を緩和してその累積時間を計測した．結果を表1に示す．キーワード総数すなわち総緩和回数は10,989回である．

語順を考慮する手法(1, 3)は考慮しない手法(2)に比べて約1.3倍の時間がかかっている．また変位を使う影響はほとんどない．1回の緩和に要する平均時間を2行目に記した．ただしここでの時間計測にはキーワードの出現位置を二次記憶から転送する時間を含めており，この時間がかなりの部分を占めている．実際に1回の緩和に要する時間は語順と変位を考慮しても1ないし2秒であり満足できる速度である．

5．2 再現率

先に述べたようにキーワード数を1まで緩和した場合，語順を考慮した場合としない場合の再現率は同じである．ただ実際にはキーワード数の大きなところで検索を打ち切るのでユーザの見る用例数は二つの手法で違ってくる．ユーザには正解だけを少数提示したい．

そこでまず語順の制約が検索数に与える影響を確認した．このため語順を考慮した手法(1, 3)としない手法それぞれについて各キーワード数での検索数を500文で合計した．結果を図2に示す．横軸はキーワード数で縦軸は対数を取った検索数である．

語順を考慮する手法の検索数はキーワード数が大きい部分で小さい．キーワードを29から18まで緩和したときの検索累計は，語順を考慮した場合が37で考慮しないと174となった．

検索数の差はキーワード数が大きい部分で顕著でありキーワード数が小さくなるに連れて小さくなりキーワード数1で逆転した．

すなわち二つの手法は同じ再現率ではあるが，語順を考慮する手法はキーワード数の大きな部分で検索結果を絞り込む効果があると言える．

5．3 適合率

手法(1, 2, 3)の検索結果の内容を評価した．

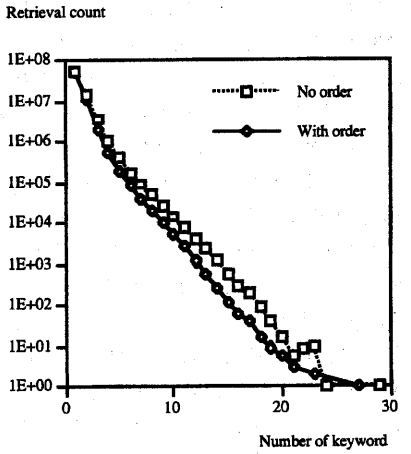

図2 各キーワードでの検索数(500文の合計)

表2 各手法の得点

手法(1)	手法(2)	手法(3)
1,593	1,075	1,273

興味があるのは上位の検索結果である．そこで先の500文から任意の58文を抽出してこれを元に現実的な入力表現を作成した．入力表現の平均文字数は26.6であった．

各入力表現の上位5つの検索結果を対象に類似性を3段階の得点(良い方から2点，1点，0点)で主観評価した．また順位を反映するため1位を5点，2位を4点…5位を1点とする重みを掛けて各手法の総得点を計算した．結果を表2に示す．これより語順と変位を使った手法が最も好ましい順序で結果を提示したことがわかる．

5．4 その他

手法(1)は入力表現のどの部分が用例のどの部分に対応したかをあいまい性なく表示できる．このため語順を考慮しない手法(2)で提示したのでは類似性が分かりにくい用例でも判定が容易になる場合がある．例えば2章の用例3の照合部分を正しく示す場合である．この特長は再現率や適合率に直接影響はしないが実際上効果的である．

6 関連研究

本研究と同じく翻訳支援を目的に類似用例を検索するシステムは多い．日英で提案されているシステムには(中村89) (隅田91) (寺濱92) (佐藤93) (武田94) (兵藤94) (北村96)などがある．こ

の中で (中村 89) は著者らと同様に入力表現と用例が共有する自立語の数に基づいた類似性計算を提案している．ただし語順を考慮しておらず，複数文節の表現を入力する実験では検索結果が「あいまい」で適合性の評価が低くなったと報告している．これは著者らの評価と一致する．

(佐藤 93) は文字を単位としてこれを連続して多く共有する文を近いと考えた「最適照合検索」を提案している．ここでは一文字を照合単位としかつ順序を考慮した照合手法 (CTM1) と二文字と三文字を照合単位とし，順序を無視した照合手法 (CTM2) を報告している．文字と単語の違いを無視すれば，著者らの語順と変位を考えた手法が CTM1 に，語順を無視した手法が CTM2 に対応する．佐藤らはこの二つの手法の適合性を評価してその結果は同等であったと報告している．ただこの評価は，日本語表現を入力して得られた上位 5 つの英訳のうち最良の英訳を使っており著者らの評価と異なっている．評価結果の違いが，評価法によるのかデータベースの性質のよるのかは今後検討したい．

(隅田 91) は構文の類似性を重要視し，助詞を検索対象に含めたシステムを提案している．このシステムは辞書の短い用例を対象にしている．著者らの用例は長いため表層の助詞の一致で構文の類似性を評価するのは難しい．さらに助詞はほぼすべての用例の複数箇所に出現するため，その位置情報を二次記憶から一次記憶に転送するだけでかなりの時間がかかる．著者らの 5.1 節と同様の予備実験では自立語を対象にする場合の約 23 倍の時間がかかることがわかっており (1 回の緩和に 26 秒かかった．自立語の場合 1.15秒であった) 現在は助詞を含めた検索はしていない．

7 おわりに

翻訳支援を目的とした類似文検索手法を提案した．この手法では入力と用例のキーワードの共有数を類似性の基準とし，その数をユーザの指示で段階的に減少させる．この時キーワードの語順とその変位を考慮しており，疑似的に構文の類似性を反映した．また入力文の形態素解析以外は表層的な手法による検索を行っており頑健である．

今後の課題を述べる．現在はすべての自立語を同等に扱った条件緩和を実行している．しかし，ユーザが知りたい表現が特定のキーワードを含む可能性は高い．そこで今後ユーザが指定するキーワードは除かないように処理を変更したい．これについては現在の動的計画法の適用改善を含めて考慮したい．

さらに，英語を始めその他の言語にも適用して多言語の用例検索システムを構築したいと考えている．これには本手法の頑健性が生かされると考えている．

【参考文献】

兵藤,池田: 係り受け構造の照合に基づく用例検索システムTWIX, 電子情報通信学会論文誌, Vol. J77-D-II (5) , pp. 1028-1030, (1994)

北村, 山本: 対訳文書の文・単語対応付け技術を利用した対訳例検索システム, 情報処理学会第53回全国大会, (2) 385-386, (1996)

熊野, 田中, 金, 浦谷: 日英ニュース原稿の対訳コーパス化に関する基礎調査, 言語処理学会第 2 回年次大会, pp. 41-44, (1996)

熊野, 田中, 浦谷, 江原: 日英放送原稿翻訳支援のための類似用例提示システム, 言語処理学会第 3 回年次大会, pp. 529-532, (1997)

Nagao, M and S. Mori: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, proc. of Coling94, Vol. 1, pp. 611-616, (1994)

中村: 用例検索翻訳支援システム, 情報処理学会第38回全国大会, (1) 357-358, (1989)

Salton, G. and M. J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, (1983)

佐藤: 用例検索による日英翻訳支援システムCTM 2--部分列インデックスを用いた最適照合検索, JAIST Research Report, IS-RR-93-6I, 北陸先端科学技術大学院大学　情報科学研究科, (1993)

隅田, 堤: 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 (10) , pp. 1437-1447, (1991)

武田, 古郡: 例文をもとにした英文書作成支援システム, 情報処理学会論文誌, Vol. 35 (1) , pp. 53-60, (1994)

寺濱, 小澤, 小嶋, 絹川: 英文作成支援における例文検索方式, 情報処理学会第45回全国大会, (3) 145-146, (1992)