要素の順序関係から見た類似文最適照合検索

市原 創　池原 悟　村上 仁一
鳥取大学大学院工学研究科
{itihara,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳の分野では用例検索による翻訳支援 [1] という考え方が提案されている。用例検索による翻訳支援システムは翻訳したい文を入力すると、それに対して表現の類似した文とその対訳をデータベースから検索する。ユーザあるいは計算機はその対訳を参照して翻訳文を作成することになる。翻訳支援を目的とした検索手法として入力文と用例文のキーワードの共有数で類似性を評価する手法や構文解析を行う手法 [2]、シソーラスを使って意味的類似性を評価する手法 [3]、そして文字に基づく最適照合検索 [4] などが提案されている。これらの手法はそれぞれに一長一短があり、それらをうまく組み合わせることで類似検索の質の向上が期待できる。

佐藤らは文字に基づく最適照合検索 [4] を提案している。具体的には、二つの文に共通に含まれる文字の数とその連続性、順序制約を考慮して類似度を計算する。これは文字に基づいた類似性評価法なので、形態素解析などの辞書的情報を利用する作業が不必要であり、頑健である。しかし、この方法では文字の照合によって予備選抜した文に対して、ダイナミックプログラミング（以下ＤＰ）法を使って再び文字列同士の照合を行っているため、十分な検索速度が得られないという問題がある。†

本論文では著者らが開発した文字単位の照合に基づく日本語類似文検索システムとそこで使用した順序関係評価法について述べる。著者らは文の類似性を判定する基準として、文字の一致数とその順序関係の一致に着目した。入力文とデータベース文の全文字を照合するために、１文字インデックスを採用した。また順序関係の一致を評価する方法としては

(A) 順序の完全一致した要素のみで評価する方法
(B) 順序関係の一致を緩やかに評価する方法

の二つを採用した。(A) は前述のＤＰ法による照合に似ているが、一致要素のポインタ (文内位置情報) 列を使って擬似的なＤＰマッチングを行っている部分が異なる。この方法だと普通にＤＰマッチングを行うより高速に順序関係の一致を評価できる。(B) は多少の順序関係の交換があっても、部分的順序関係の一致を評価できるように、順序一致率という評価値を考案し用いた。本システムは7Mバイトの日本語テキストデータベースに対して、平均約0.6秒の検索速度を達成しているので十分実用的であると考える。

2 文字の順序関係から見た日本語文の類似度

2.1 基本ヒューリスティックス

ここでは表層的な文字によって日本語文の類似性を判定する基本的なヒューリスティックスとして以下の２つの仮定を採用する。

仮定１：一致する文字ができるだけ多い文字列は似ている。

この仮定は文字による類似性評価では必要条件である。なぜなら最も似ている文（まったく同じ文）は、全ての文字が一致する文であり、最も似ていない文は一つも文字が一致しない文だからである。

仮定２：一致文字の順序関係ができるだけ保たれている文字列は似ている。

順序関係の評価については、例として入力文に対して次の３つの用例が与えられる場合を考えてみる。

入力文：私は問題を発見した。
用例Ａ：彼は法則を発見した。

†その後、佐藤らは部分文字列の一致数だけで類似性を評価する最適照合検索を提案している [5]。この方法は部分文字列に基づく類似性評価を前提としており、単語単位の照合の場合はまた別の評価が必要であると考える。著者は将来的に単語単位の検索との比較をねらっているため、その拡張の容易性を考慮して、１文字の照合に基づくシステムを開発した。

用例Ｂ： 発題したの は私だ。
用例Ｃ：その法則 を 彼 は発見した。

文字の一致部分はわかりやすくするため、下線で表した。用例Ａと用例Ｂを単に一致文字数のみで比較した場合、どちらとも７個で、同じ評価となる。しかし、我々の感覚では用例Ｂよりも用例Ａの方が似ていると感じるだろう。なぜなら用例Ａの場合、一致部分が「～は～を発見した。」のようにある言い回しになっており、文の構造的な類似が見られるからである。このように文の類似性評価では、文構造の一致を評価することが重要となる。特に翻訳支援では離散型共起表現 [6] のような使用頻度の高い表現や固定的な言い回しなどの表現の含まれる文を抽出したい。

このような類似文を発見するには、構文解析をする方法も考えられるが、ここではもっと頑健で高速な表層的手法を採用する。すなわち、一致した文字の出現順序の一致を調べることで構造的類似性を近似的に評価する手法である。このような評価法を以下では一致文字の順序関係評価法と呼ぶ。

では、今度は用例Ｃのような場合はどうだろうか。この場合一致文字は用例Ａとまったく同じだが、格助詞「は」と「を」の部分で順序の交換があるのがわかる。にもかかわらず文構造の類似性という観点から見れば、用例Ｃも用例Ａと同様に、文構造の類似が見られ、似ていると言える。このように日本語は文要素の順序交換が比較的自由な言語であるため、このような部分的順序の交換がありうることも考慮しなくてはならない。

2.2 順序関係評価法

以上のようなヒューリスティックスに基づいた日本語文類似性判定のために、次の２種類の一致文字順序関係評価法を採用した。

1. 順序の完全一致した文字のみを評価する方法。
2. 多少の順序の交換があっても部分的順序一致を評価する方法。

1,2 の評価手法を以下では、それぞれ順序完全一致評価法、順序一致率評価法と呼ぶ。また、これらを用いた検索をそれぞれ順序完全一致検索、順序一致率検索と呼ぶこととする。

順序完全一致評価法

	や	ま	が	た	か	い
た				1		
か					2	
い						3
や	1					
ま		2				

図 1: 文字一致マトリックス

順序完全一致評価法は、照合で一致した文字の中でも順序の完全一致した文字のみを一致文字数として評価の対象とする方法である。例えば、次のような簡単な例を考えてみる。

A = やまがたかい
B = たかいやま

二つの文字列 A, B の間には (や・ま・た・か・い) の文字が一致している。文字の一致をマトリックス状に表すと、図１のようになる。これを見ると、順序の完全に一致している組み合わせは (やま) と (たかい) の２つがあることがわかる。順序完全一致評価法ではこの２つの組み合わせのうち一方を選んで類似度として評価し、一方を無視する。通常、評価する一致文字は多い程よいので、この場合は３文字である (たかい) を選択する。このように順序の一致した文字列を照合する方法としてはＤＰ法が一般的である。しかし、ＤＰ法ではマトリックス中の一致していない部分の計算も含まれるため、無駄が多い。

提案手法ではＤＰ法で必要のない計算を省くために、直接文字列を照合せずに、一致文字のポインタ（文内位置情報）を使って最大一致数を計算する。例えば文字列 A, B に次のようなポインタ番号が付加されているとする。

A = や$_1$ま$_2$が$_3$た$_4$か$_5$い$_6$
B = た$_1$か$_2$い$_3$や$_4$ま$_5$

すると一致文字の集合 (や, ま, た, か, い) は $[(1,4)(2,5)(3,1)(4,2)(5,3)]$ のようなポインタ対の集合に書き直すことができる。これを文字列 A, B 間で n 個の一致がある場合に一般化する。

$$A = a_1 a_2 ... a_{l_A}, B = b_1 b_2 ... b_{l_B} \quad (1)$$

$$[(ap_1, bp_1)(ap_2, bp_2)...(ap_n, bp_n)] \quad (2)$$

これらのポインタ対は文字列Ａのポインタ番号順に並んでいる必要がある。すると２文間の最大一致数

onm と類似度 $sim1$ は次のような式で計算できる。

$$sim1(A,B) = \frac{onm}{l_A} \times \frac{onm}{l_B} \tag{3}$$

$$onm = \begin{cases} 0 & n = 0 \\ 1 & n = 1 \\ \max_{1 \le i \le n} ms_i & n > 1 \end{cases} \tag{4}$$

$$ms_i = \begin{cases} 1 & i = 1 \\ 1 + \max_{1 \le j \le i-1} s_{i,j} ms_j & i > 1 \end{cases} \tag{5}$$

$$s_{i,j} = \begin{cases} 1 & ap_i > ap_j \wedge bp_i > bp_j \\ 0 & otherwise \end{cases} \tag{6}$$

順序一致率評価法

順序一致率評価法では多少の順序の交換があっても部分的な順序一致を評価するために順序一致率という評価値を考案した。順序一致率の計算も式 (2) のポインタ列を使用する。式 (1)(2) が定義されるとき、順序一致率 ord と類似度 $sim2$ は以下のように定義する。†

$$sim2(A,B) = \frac{nm}{l_A} \times \frac{nm}{l_B} \times ord \tag{7}$$

$$nm = \sum_{j=1}^{l_B} \sum_{i=1}^{l_A} m_{i,j} \tag{8}$$

$$m_{i,j} = \begin{cases} 1 & a_i = b_j \\ 0 & a_i \neq b_j \end{cases} \tag{9}$$

$$ord = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} o_{i,j}}{\frac{n(n-1)}{2}} \tag{10}$$

$$o_{i,j} = \begin{cases} 1 & (ap_i > ap_j \wedge bp_i > bp_j) \\ & \vee (ap_i < ap_j \wedge bp_i < bp_j) \\ 0 & otherwise \end{cases} \tag{11}$$

3 類似文検索システム

ここでは著者らが開発した類似文最適照合検索システムについて説明する。検索対象とするテキストデータベースは、1995 年毎日新聞経済記事 1 年分（78842 文、7Mbyte）を利用した。文字に基づく最適照合検索では入力文とデータベース中の文の全ての文字を照合する必要がある。これを高速に行うためにあらかじめデータベースを加工して 1 文字インデックスを作成した。これは n=1 の N-gram インデックス [7] に相当し、各文字に文番号とポインタが対応している。図 2 にインデックス作成の例を示す。

テキストデータベース

テキスト	た	か	い	や	ま	。	か	た	い	い	た	。
文番号	1	1	1	1	1	1	2	2	2	2	2	2
ポインタ	1	2	3	4	5	6	1	2	3	4	5	6

↓ 文字コード順にソート

1 文字インデックス

文字	文番号	ポインタ
。	1,2	6,6
い	1,2,2	3,3,4
か	1,2	2,1
た	1,2,2	1,2,5
ま	1	5
や	1	4

図 2: 1 文字インデックス作成例

全体的な類似文検索は以下の手順で実行される。まず入力文が与えられると、入力文の各文字を文字コード順にインデックスと照合する。照合は一回のサーチでよいので高速に一致文字の文番号とポインタを得ることができる。検索システムはまず文番号によって各文ごとの文字一致数（式 (8) の nm に相当）をカウントし、文字一致率 rmc を求める。

$$rmc(A,B) = \frac{nm}{l_A} \times \frac{nm}{l_B} \tag{12}$$

文字一致率が求まると検索対象文を文字一致率の高い順に全データベース文の数%に絞り込む。最後に絞り込まれた文に対して、ポインタ情報を使って順序関係評価と類似度計算を行い、類似度の高い順にランキングして類似文として出力する。図 3 に全体的な類似文検索の手順を示す。

4 検索実験と考察

検索システムを使って、次の 3 つの手法を対象に検索時間の計測と検索結果の内容評価を行った。

(1) 文字一致率のみの評価による検索
(2) 順序完全一致検索
(3) 順序一致率検索

入力文は 1994 年毎日新聞経済記事から任意に選択した。検索時間の計測は入力文 5000 文を用いて、検索時間の総和平均を求めた。表 1 に検索時間の計

†$nm, o_{i,j}$ は厳密には、1 文字に対して複数の一致がある場合はカウントしない。ここではその処理は省略し簡単に式 (9)(11) のように定義した。

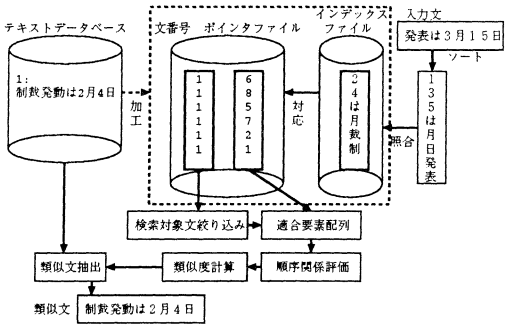

図 3: 検索アルゴリズムの概要

測結果を示す。順序完全一致検索と順序一致率検索では平均0.6秒という実用的な検索速度を達成している。検索内容の評価は入力文100文を用いて、検索結果に3段階の得点(2点、1点、0点)をつけて主観評価した。† 各手法の得点合計を表2に示す。2点のついた文は手法(1)で全体の1割、手法(2)で2割強、手法(3)で2割弱程度あった。これは類似度0.2以上の文は全体の4分の1程度であることを考えると、妥当な結果である。

文字一致率評価のみの手法(1)は検索時間はかなり高速だが、得点で比較すると最も低く、順序関係評価の効果がわかる。手法(2)と手法(3)は検索速度も得点もほぼ等しいことがわかった。しかしお互いの検索結果が一致しているのは4割程度である。結果には次のような順序の交換を許すか許さないかの違いも表れている。

入力文 年功序列、終身雇用は、おのずと崩れる。
手法 (2) また、「年功序列、終身雇用は過去のものとなったと考えてほしい。
手法 (3) 終身雇用、年功序列が崩れている。

表 1: 検索時間 (ms)

手法	(1)	(2)	(3)
平均検索時間	170	572	538

表 2: 各手法の得点合計

手法	(1)	(2)	(3)
得点合計	57	76	71

†得点は定型表現などの一致の多く見られるものを高く評価した。

5 おわりに

本論文では、文字の一致と順序関係から見た最適照合検索手法を提案した。この手法では文の構造的類似性を近似的に判定するために、2種類の順序関係評価法を採用した。順序関係評価の計算は一致文字のポインタ情報を使うのでDP法より高速である。また、文字を照合単位としているため助詞なども含めた表現を発見できる。開発した検索システムで検索実験を行ったところ実用的な検索速度を得た。

この手法は単語に基づく最適照合検索との比較をねらって考案した。本手法は単語を照合単位とした場合にも容易に拡張できる。今後は検索システムを単語単位照合へ拡張し、文字照合との比較を目指す。

また、用例検索による翻訳支援では不一致部分の単語等の規則性を明確にしたり、その置き換えを行ったりするのも重要な技術である。今後、その可能性をさぐる研究も進める必要がある。

参考文献

[1] 隅田, 堤:翻訳支援のための類似用例の実用的検索法, 電子通信情報学会論文誌,D-II, Vol. 74-D-II, No. 10, p. 1447 (1991).

[2] 兵藤, 河田, 応, 池田:構文付きコーパスの作成と類似用例検索システムへの応用, 自然言語処理, Vol. 3, No. 2, pp. 73–88 (1996).

[3] 大井, 隅田, 飯田:単語間の意味的類似度に基づく文書検索手法, 言語処理学会第2回年次大会発表論文集, pp. 109–112 (1996).

[4] Sato, S.: CTM: An Example-Based Translation Aid System, Proc. of COLING-92, Vol.IV, pp. 1259–1263 (1992).

[5] 佐藤:用例検索による日英翻訳支援システムCTM2 -部分列インデックスを用いた最適照合検索-, JAIST Reserch 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).

[6] 池原, 白井, 河岡:大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会, Vol. 36, No. 11, pp. 2584–2596 (1995).

[7] 長尾編:自然言語処理, 岩波講座ソフトウェア科学, 433 pp. (1996).