5Y-02　日本語の構造的類似文検索アルゴリズムの検討

谷口雄介　池原悟　村上仁一
鳥取大学大学院工学研究科

1. はじめに

機械翻訳の分野では、翻訳精度を向上させるため、用例を利用する方法が提案されている。用例翻訳に利用される用例検索システムは、翻訳する文に対して最も表現の類似した文とその対訳を DB(データベース) から高速に検索することが求められる。

これに対し本研究では、日本文の持つ係り受け関係を利用する。文節間の係り受けにより構造の類似性を判定し、入力文と同じパターンの係り受け関係を持つ類似文を検索する手法を提案する。

同様な係り受け関係を利用した検索システムが、論文 [1] で提案されている。[1] は DB から指定した係り受け関係の全てを含む文を検索する手法である。これに対して本研究では、入力文と一致する係り受け関係を最も多く含む文を類似文として検索の対象とする。

2．係り受け関係から見た文の構造的類似性

2.1 日本文における係り受けの特徴

類似文の決定において、本研究では以下の係り受けの特徴を利用する。

- 係り受け関係は互いに交差しない
- 係り先は後方のみである
- 係り先は 1 つだけである

2.2 類似性判定アルゴリズム

入力文に対して、DB 文中から文節間の係り受け関係が最も類似した文を抽出するには、入力文と同一種類の文節 (以下一致文節と呼ぶ) が含まれている必要がある。一致文節を含む DB 文に対して、係り受けを構成する文節を 1 つのペアとして扱い、「係り受け関係の一致する文節ペア数が最大の文が構造的に類似した文である」とし、DB から最適な文を決定する。アルゴリズム全体の流れを図 1 に示す。

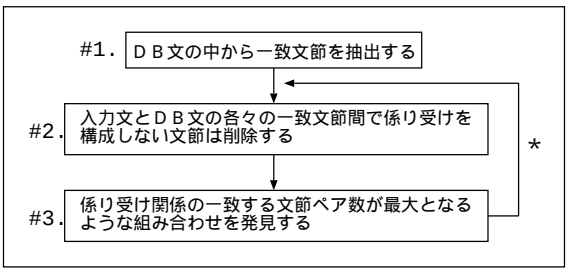

図 1. アルゴリズム全体の流れ

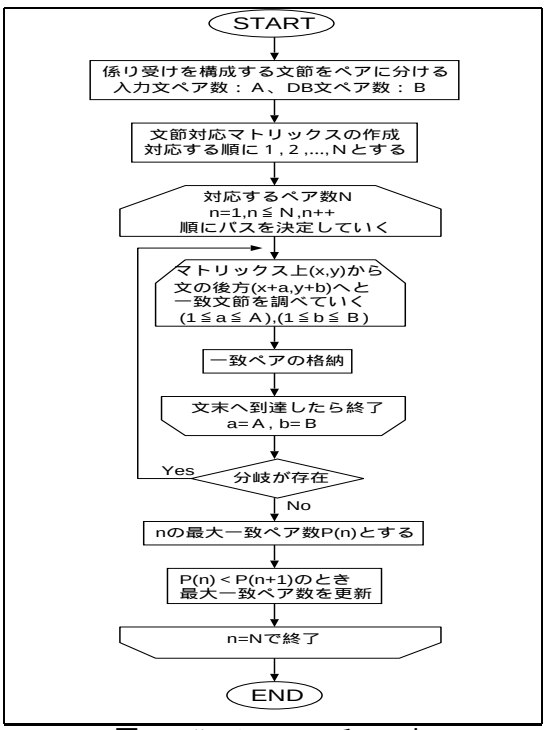

図 2．#3 のフローチャート

#1 により DB から一致文節を含む文を抽出する。抽出された文を一致文節数が多いものから順にソートする。上位の候補から順に#2、#3 を適用し、各候補に対して係り受けの一致 (一致ペア数) が最も多くとれるパターンを決定する。このとき得られた一致ペア数から、繰り返し条件を設定する。得られた一致ペア数を n とし、各候補文の一致文節数を m とする。条件 $m > n$ を満たす候補に対して#2、#3 を適用する。最終的に、得られた一致ペア数の中で最大の候補が最も類似した文として決定される。次に#3 を説明する。

Retrieval Algorithm for Structurally Similar Sentences
Yusuke TANIGUCHI,Satoru IKEHARA,Jin'ichi MURAKAMI
Tottori University

2.3 同一係り受け文節ペアの発見

2.3.1 係り受け文節ペアの発見

一致ペアの発見手順 (#3) をフローチャート (図 2) で示す。候補文と入力文との対応を調べ、各々について最大一致ペア数を決定する。

2.3.2 一致ペア発見の例

図 3 の例文 1 に 2.3.1 で述べた#3 を適用する。図 3 中の記号を以下に示す。

- A ~ D：文節の種類 (同じアルファベットは一致文節)
- 矢印：文節間の係り受け関係
- 文節下の数字 (1 ~ 5)：文節
- (数)：入力文の係り受けペア番号
- [数]：DB 文の係り受けペア番号

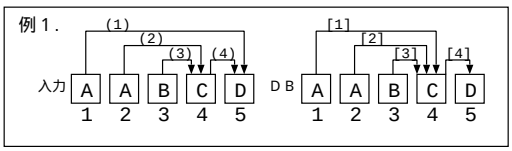

図 3 . 例文 1

入力文側の係り受けペアに対して、DB 文側の係り受けペアの対応を示す文節対応マトリックス (図 4) を作成する。図中の○ は対応する文節が存在することを示す。

DB文 / 入力文	[1] AC 14	[2] AC 24	[3] BC 34	[4] CD 45
(1) AD 15				
(2) AC 24	α	β		
(3) BC 34			γ	
(4) CD 45				δ

図 4 . 文節対応マトリックス

文節対応マトリックスを右下方向 (文の後方) へと進み、対応する文節の存在を調べていくことで一致文節ペア最大のパスを決定する。図 4 においては、矢印で結んだパス (α,γ,δ) と (β,γ,δ) が最も多く対応がとれ、一致ペア数 3 が決定される。

2.3.2 係り受け不一致の問題

図 2 では、入力文の文節に対して DB 文の文節が 1 対 1 に対応せず構造に違いが生じることがある (図 5)。この問題は、入力文と DB 文の文節番号を照合することで解決できる。

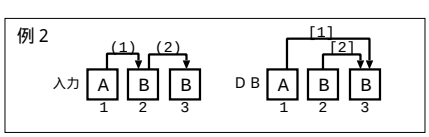

図 5 . 構造的に矛盾するペア

3 . 評価実験

3.1 アルゴリズム動作の確認

提案したアルゴリズムの動作を確認するため、係り受けを独自に与えた DB を作成し実験を行う。実験により各入力文に対して類似文を決定することができ、他の文節をまたいで係るような複雑な係り受けパターンに対しても、適用できることが確認できた。

3.2 性能評価実験

実際にコーパスに対してアルゴリズムを適用し、その性能を評価する。検索対象データベースには京都大学テキストコーパス (毎日新聞記事約 2 万文) を使用し、各文節の品詞並びが同一のものを一致文節とする。入力文は京大コーパスの中から 30 文程度を選び、入力文・DB ともにあらかじめ文節を記号で置き換えて用いる。実験の結果を表 1 に示す。表中の抽出数は#1 により DB から抽出された一致文節を含む文の数、適用数は繰り返し回数 (図 1 中の *) を表わす。最大ペアは得られた一致ペア数の中で最大のものとその件数を表わす。検索時間は DB から一致文節を抽出して各候補の一致ペア数を決定し、条件を満たして終了するまでにかかった時間を表わす。

表 1 より、文節数の増加に伴い抽出数・適用数が多くなり検索時間が増加するという結果が得られた。

表 1 . 実験結果

文節数	抽出数	適用数	最大ペア	検索時間 (s)
5	9391	1112	4(2)	1.29
8	9122	1280	4(6)	1.46
14	12603	1433	6(1)	1.69
26	14054	2933	5(19)	3.67

3.3 考察

検索時間を抑えるためには、適用数を入力で指定する方法や、3 つの手順を繰り返すアルゴリズム自体の見直しが必要となる。また、文節の種類を「名詞・助詞」のように大きく分類すると得られる結果が多くなる。結果を絞り込むため、一致文節をより詳細に分類するなどの方法が必要となる。

4 . 今後の課題

本研究では、文節間の係り受け関係を利用した類似文検索アルゴリズムを提案した。実験によりアルゴリズムの動作も確認され、複雑な係り受け関係を持つ文も類似文を決定することができた。今後は、高速化も含めたアルゴリズムの見直し、文節の種類を決定する変換テーブルの作成などが挙げられる。

参考文献

[1] 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, 自然言語処理学会論文誌 Vol.3,No.2(1996).

[2] 市原, 池原, 村上：要素の順序関係から見た類似文最適照合検索, 自然言語処理学会第 5 回年次大会 (1999).