E-002 構造的類似文検索アルゴリズムを応用した日本語文型パターン抽出法
Sentence pattern extraction using structural similarity search algorithm

田中 康仁† 村上 仁一† 徳久 雅人† 池原 悟†
Yasuhito TANAKA Jin'ichi MURAKAMI Masato TOKUHISA Satoru IKEHARA

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ある。従来の翻訳に利用される用例検索システムでは、品詞の並びのみで文と文の類似度を調べる。しかし、単に品詞の並びだけを調べるだけでは、文の木構造が一致するとは限らず、不適切な用例を検索してしまうという問題が発生する。この問題を避けるために、係り受け関係を利用する方法 [1] が提案されている。

係り受け関係を用いた類似文検索アルゴリズムに「依存構造を考慮した文型パターン検索アルゴリズム」[2] がある。このアルゴリズムは、入力文と検索の対象となるデータベース文の間で、もっとも係り受け関係の近い文を抽出できる。

本研究では、このアルゴリズムを用いて、入力文に対する類似文をデータベースから抽出する。さらに、抽出された文と入力文の対訳をみて、対訳文における文法構造の類似性を調査し、係り受け関係が用例翻訳に有効であるか検証する。

2．係り受けを用いた類似文の抽出

2.1 使用データベース

検索対象データベース (DB) に、日英対訳例文集 [3](約 8 万文) を用いる。係り受け解析情報を得るために、日英翻訳システム ALT-J/E(NTT) を使用する。品詞情報についても、ALT-J/E の品詞コードを利用する。

2.2 文節の分類

類似文検索アルゴリズムは、文中の文節単位の係り受けの一致をもとに、類似文を検索する。アルゴリズムを使用するには、検索の判断基準となる文節の種別を決定する必要がある。文節の種別は、「品詞の並び (品詞列)」で定義する。

本研究では、アルゴリズムで用いる品詞の種類を 18 種とした。DB 文中に現れる文節の種類 (品詞種の組合せ) は、1,738 種であった。文節の例を表 1 に示す。

表 1. 文節の例

品詞列	例
名詞+格助詞	娘を，語尾を
動詞+接続助詞	集めて，見合わせて
接頭辞+名詞+格助詞	ご挨拶を，ご要望に

2.3 日本語類似文抽出実験

2.2 節で述べた品詞分類をもとに日本語の類似文抽出実験を行う。入力文は例文集の中から無作為に選んだ 100 文である。各入力文ごとに、DB の 8 万文を対象とした抽出実験を行う。

†鳥取大学工学部 知能情報工学科

類似文の抽出例を図 1 に示す。図 1 のように、1 つの入力文に対し、複数の抽出文 (候補文) が出力される場合がある。

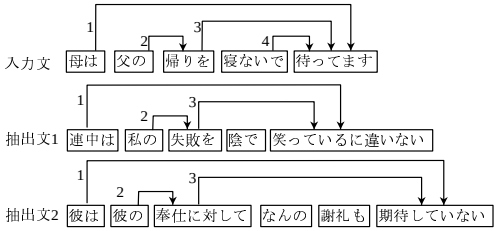

図 1: 類似文の抽出例

実験の結果を表 2 に示す。縦軸の見出しは、入力文の係り受け関係の文節ペア数を表す。見出しの括弧内の数字は、入力文の数を表す。横軸の見出しは、DB 文との間で一致した係り受け関係の文節ペア数を表す。表内の上段の数字は、候補文が抽出された入力文の数、下段の (数値) は抽出された候補文の総数を表す。「-」は候補文が未抽出の場合を表す。

図 1 の「入力文」の文節ペア数は 4 であり、対する「抽出文 1」との一致文節ペアは 3 である。したがってこの「入力文」は、表 2 中の横軸 3 と縦軸 4 の交点 8 文の中の 1 文である。

表 2. 実験結果

	6	5	4	3	2	1	0
10(3)	- -	- -	2 (9)	1 (11)	- -	- -	- -
9(3)	- -	- -	3 (6)	- -	- -	- -	- -
7(2)	- -	- -	1 (3)	- -	- -	- -	1 -
6(11)	1 (1)	3 (17)	4 (13)	2 (41)	1 (19)	- -	- -
5(25)	- -	7 (69)	4 (9)	9 (170)	4 (144)	- -	1 1
4(28)	- -	- -	12 (318)	8 (83)	7 (430)	1 (10)	- -
3(16)	- -	- -	- -	11 (580)	5 (94)	- -	- -
2(10)	- -	- -	- -	- -	8 (44)	2 (175)	- -
1(2)	- -	- -	- -	- -	- -	1 (125)	1 (1)

実験の結果、アルゴリズムは入力文 100 文中 97 文に対し、候補文を出力した。

3．対訳の類似性の判定

日本語の入力文と候補文に構造的類似性のあるそれぞれの対訳は、類似性が存在すれば、用例翻訳に有効であると考えられる。

本節では、2.3 節の実験で得られた抽出文と入力文について、それぞれの対訳文に類似性があるかを、人手により調べる。

入力文と抽出されたDB文の係り受け関係が一致している部分について、対訳の主語句 (S)、述語句 (V)、補語句 (C)、目的語句 (O) の文法構造が同じであれば、類似していると判定する。ただし、ある入力文に対し複数のDB文が抽出されている場合は、抽出文中の1文以上に、対訳の文法構造の一致がみられれば、類似しているとする。

対訳が類似していると判定した入力文と抽出文の例を図2に示す。

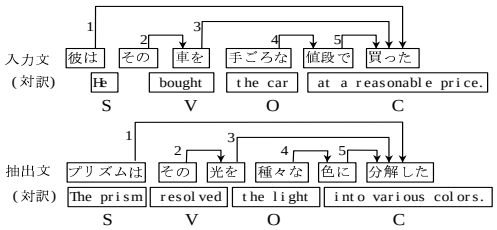

図 2: 対訳に類似性のある例

表3に結果を示す。表内の数字は、類似していると判定した文の数を表す。

表 3. 対訳類似性の判定結果

	6	5	4	3	2	1	0
10	-	-	0	0	-	-	-
9	-	-	0	-	-	-	-
7	-	-	1	-	-	-	0
6	0	0	2	1	0	-	-
5	-	3	0	2	1	-	0
4	-	-	5	2	4	0	-
3	-	-	-	7	0	-	-
2	-	-	-	-	3	1	-
1	-	-	-	-	-	1	0

日本語類似文 (候補文) が存在する97文の入力文中、33文に類似性ありと判定した。

4．考察

日本語類似文抽出実験の結果、入力文100文中、類似文検索アルゴリズムが候補文を抽出した入力文は97文であった。また対訳類似性の判定で、類似性ありと判定した文は、97文中33文であった。

4.1 抽出方法の問題

日本語類似文抽出実験では、より多くの候補文を抽出する目的で、品詞の種類を18種と少なめに定義して実験を行ったため、殆どの入力文に、候補文が抽出されたと考えられる。反面、不適切な候補文もともに抽出してしまうため、適切な候補文を漏らさない程度の詳細な品詞分類が望まれる。

一方、入力文に対し適切な候補文であるにも関わらず、DB文から抽出されない文があった。これは、文節内に複合語が含まれている場合、複合語としての品詞種が同様でも、複合語内の品詞並びが異なるため、文節種が異なると判定され、文が抽出されなかったと考えられる。表4に複合語の例を、図3に複合語が未考慮のため、抽出漏れを起こした例を示す。

表 4. 複合語の例 (複合名詞)

字面	損害/補償/を		嘔吐/を
文節	(名詞)+格助詞	=	(名詞)+格助詞
品詞列	名詞+名詞+格助詞	≠	名詞+格助詞

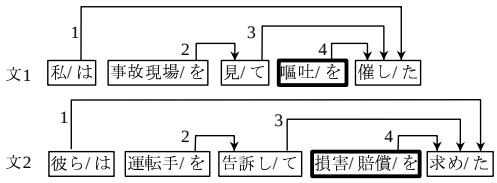

図 3: 複合語の例

4.2 対訳類似性の判定

類似性ありと判定した対訳の中に、特殊な構造を持つもの (図4) があった。本実験において、データベースの対訳の中に、対応する特定の構文が存在したため、類似性があると判断できた。

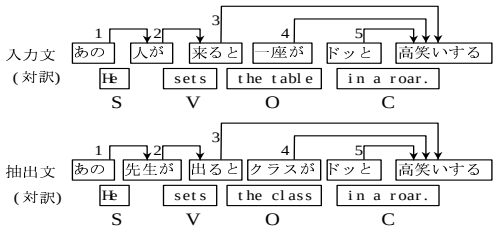

図 4: 対訳に類似性のある例2

また、類似性がないと判定した対訳の中に、日本文に現れない主語のため対訳の構造が異なるもの (図5) があった。これは、主語を省略する傾向の強い日本語と、省略しない英語の特性を現している。

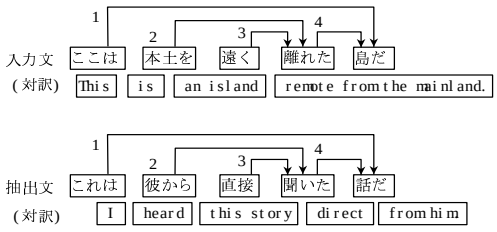

図 5: 対訳に類似性のない例

5．まとめ

本研究では、日本文の係り受け関係を利用した手法が、翻訳の用例検索をおこなう際、有効であるかを検討した。実験の結果、本手法で抽出した候補文に、適切な対訳が存在することが確認できた。しかし、候補文が抽出された入力文に対し、候補文の対訳の中に適切な文が存在した比率は約1/3であった。

今後、日本語類似文の検索精度の向上を目指す。検索精度の向上には、不適切な候補文の抽出と適切な候補文の抽出漏れを防ぐ必要がある。

不適切な候補文の抽出を抑えるために、「品種の種別の細分化」「候補文判定条件へ係り受け種別の追加」を考えている。また、適切な候補文の抽出漏れを防ぐために、「複合語の単品詞 (句) 化」などを行って行きたい。

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, pp.73-88(1996).
[2] 谷口, 池原, 村上：構造的類似文検索アルゴリズムの検討, 情報処理学会第63回全国大会, Vol.2, pp.247-248(2001).
[3] 村上, 池原, 徳久：日本語英語の文対応の対訳データベースの作成, 「言語、認識、表現」第7回年次研究会 (2002).