コーパス構造化への類推関係抽出手法

イヴ・ルパージュ & 中岩 浩巳
{yves.lepage, hiromi.nakaiwa}@atr.co.jp
エイ・ティ・アール - 音声言語コミュニケーション研究所

1 はじめに

多くの自動翻訳システムでは、コーパスを分析し、人手でルールやパターンを作成する。また、ルールやパターンを効果的に作成するためには、コーパスの中の言語現象を自動的に分類構造化する前処理が有効である。一般に、これは人手により行なわれているため、時間的・金銭的コストが問題となる。よって、この自動化が望まれている。本研究では、普遍性を考慮しつつ、言語的な現象に基づいて、前処理を支援する手法を提案する。本手法では、コーパスにある類推関係の抽出をコーパスの構造化への一段として試みた。

2 コーパスからの類推関係抽出

2.1 類推関係の定義

類推関係というのは、例えば、

もちろんだよ。 : この近くだよ。 ::
もちろんです。 : この近くです。

の様なものである。言語学でよく知られている現象であり、次の様に定式化できる (Lepage 01)。

$$A : B :: C : D \quad \Rightarrow \quad \begin{cases} \mathrm{dist}(A,B) = \mathrm{dist}(C,D) & (1) \\ |A|_a + |D|_a = |B|_a + |C|_a & (2) \end{cases}$$

ここでは、a は文字を示し、$A \sim D$ は表現である。文字 a が表現 A 中に出現する頻度を $|A|_a$ で表す。また、$\mathrm{dist}(A,B)$ は A と B の間で削除を 1・挿入を 1 のコストで計算した編集距離である。(1) を編集距離制約と呼び、(2) を文字制約と呼ぶ。

2.2 表現の間の類推の計算

編集距離制約 (1) に含まれている編集距離を計算するためには、一般的に自動計画アルゴリズム (DP matching) が用いられる (Wagner & Fischer 74)。しかし、ここでの編集距離 $\mathrm{dist}(A,B)$ と文字列類似 $\mathrm{sim}(A,B)$ の間は以下の式が満たされることが知られている。

$$\mathrm{dist}(A,B) = |A| + |B| - 2 \times \mathrm{sim}(A,B)$$

ここで、$|A|$は表現 A の長さを表す。このうち、文字列類似の計算は 2 進数記号化を用いることにより速く計算できる (Allison & Dix 86)。さらに、文字制約 (2) も 2 進数記号化し、ビットを数える操作により速く計算できる。計算時間を比較したところ、表 1のとおり 2 進数記号化の効果は大きく、本研究では 2 進数記号化手法を採用することにした。

表 1: 類推関係の計算時間 (Pentium4-2 GHz, メモリ 1Gb ; 一万の類推関係の 1 件当たり : μs)

手法	最低	平均	±	標準偏差	最大
2 進数	1	2.26	±	1.06	13
自動計画	20	43.53	±	33.18	1028

2.3 表現の組み合わせの成立

コーパスから上記の定義を満たす 4 つの表現の組み合わせを抽出すると無意味な出力が多く出現する危険がある。例えば、理論的にいずれの A と B の表現でも次の類推関係を満たす。

$$A : A :: A : A \qquad A : A :: B : B$$

しかし、そのような類推関係は自明であり、その抽出は無意味である。また、 $A : B :: B : C$ のような類推関係も可能だが、類推関係は変換ルールに近いものであると考えると、そのような関係は言語学的に適切ではない。

すぐできます : すぐにできます :: すぐにできます : x
$\Rightarrow$ $x =$ *すぐににできます ?

最後に、1 つの類推関係には 8 種類の異なる書き方があるので、実際上、その内の一種類を抽出するだけで充分である。

$$\begin{aligned} A : B :: C : D \\ &\Leftrightarrow A : C :: B : D \\ &\Leftrightarrow B : A :: D : C \\ &\Leftrightarrow B : D :: A : C \\ &\Leftrightarrow \ldots \end{aligned}$$

以上の制限を用いて、必要で充分な類推関係だけを抽出する方法を提案する。

単純に4つの表現の組み合わせをすべて検討する場合、N 表現のコーパスに対して $O(N^4)$ の時間がかかる。しかし、3つの表現を固定し、4つ目の表現を推定する類推方程式を用いると $O(N^3)$ となる(Lepage & 白井 02)。また、文字制約(2)を用いて、すべての表現 A と B の組み合わせに対し、文字出現頻度差 ($|A|_a - |B|_a$) を計算し ($O(N^2)$) 、出現回数が2以上文字出現頻度差を記憶する。そして、記憶した頻度差を持つ組み合わせについて、CD を計算することで最悪でも $O(N^3)$ の計算量での抽出が可能となる。以上の提案した方法で、抽出時間を計った結果を表2に示す。

表 2: 類推関係の数と抽出時間 (CPU は表1と同じ)

文の数	類推関係の数	抽出時間 (時: 分: 秒)
2 500	70	2:48
5 000	479	54:42
10 000	4984	7:51:51

3 抽出した類推関係の分析

類推関係の分類を検討するため一部のデータで分析を行なった。ATRの「旅行会話基本表現集」コーパスの最初の1000文を対象に、類推関係を抽出した結果、23個抽出することができた (付録参考) 。それらについて以下の分類を試みた。

3.1 形態論：活用

9a. ~ 9b. では現在・過去の対立が見られる。類推関係で活用や曲用を簡単に容易に抽出することができる。類推という現象は普遍的であることから、他の言語でも活用や曲用のような範列論的表現対が検索できる。例えば、豊かな曲用を持つポーランド語のコーパスを使用した場合、主格・対格・属格等の表現対が抽出できた (下記左) 。また、動詞の意味を変換する接頭辞も抽出できた (下記右) 。

agencja : decyzja	rzucili : wrzucili
agencję : decyzję	łączyli : włączyli
agencji : decyzji	szedłem : wszedłem
...	...

英語では、単数・複数の対立や -ed または -ing の接尾辞で作られる過去動詞形または現在分詞が抽出される。また、より興味深いことには、派生形態の現象も抽出される。

analyzed : analyzer	classification : classificatory
assembled : assembler	compensation : compensatory
...	...

3.2 語彙：単語クラス

1a. ~ 4b. では「いつ・いくら・どちら・なぜ」の入れ換え現象が抽出された。これらは、アメリカの構造主義派の手法のように、その入れ換え可能性に基づいてクラスに分類でき、疑問詞のクラスに入る。抽出された類推関係は、付録の形式で出力されているが、下記の形式の方がよりコンパクトで現象の理解が容易である。

いつ。 : いつですか。
いくら。 : いくらですか。
どちら。 : どちらですか。
なぜ。 : なぜですか。

すなわち、

考察1：コーパスの構造化のためには、一覧で検討できる形式が望まれる。

コーパス全体から「いつ・いくら・どちら・なぜ」を含む文を用いることで、その疑問詞についてより詳細な検討ができる。例えば、以上の四つの疑問詞の中からは、「いつ」だけが前・後置が可能であることが類推関係により明らかになった。

いつお戻りですか。 : お戻りはいつですか。
いつ開くのですか。 : 開くのはいつですか。
いつご出発ですか。 : ご出発はいつですか。

「いつ」以外では、例えば「一泊はいくらですか。」から作られた文「* いくら一泊ですか。」は非文である。

6a. ~ 6b. では、意味的な対立「大きい・小さい」に基づいた置換が見られる。しかし、本コーパス全体では、その現象に限らず、「左・右」「夫・妻」等の極的な対立も見られる。

その二つ右です。 : その二つ左です。
あちらの右側です。 : あちらの左側です。
あの信号を右折です。 : あの信号を左折です。
...

妻と一緒です。 : こちらは私の妻です。
夫と一緒です。 : こちらは私の夫です。

同じように、17a. ~ 18b. では、「日本語・英語」の入れ換えが起こっており、その1000文だけを見るとで極的な対立のように見える。

以上のような「X・Y」の極的な対立を見ると、「X」→「Y」の変換ルールが成立可能であると考えられる。例えば、「小さ」→「大き」の変換ルールが想像できる。すなわち、全コーパス中にある「あなたの声は小さい。」と言う文に対して、コーパスに存在しない「あなたの声は大きい。」と言う文が欠けているという結論が導きだせる。しかし、変換ルールだけを考えると、生成しすぎる危険性がある。例えば、「あなたは大都市と小さな町では、どちらに住むのがより好きですか。」→「あなたは大都市と大きな

表 3: 類推関係行列

これいくら。	それいくら。	6/35	一日いくら。	一泊いくら。	10/35
これはいくら。	4/35	2/35	2/35	4/35	2/35
(a) 12/35	それおいくら。	5/35	一日おいくら。	一泊おいくら。	一晩おいくら。
これいくらですか。	それいくらですか。	7/35	一日いくらですか。	一泊いくらですか。	一晩いくらですか。
これ、いくらですか。	それ、いくらですか。	あれ、いくらですか。	12/35	一泊、いくらですか。	一晩、いくらですか。
これはいくらですか。	それはいくらですか。	あれはいくらですか。	10/35	一泊はいくらですか。	9/35
6/35	8/35	(b) 2/35	一日でいくらですか。	一泊でいくらですか。	5/35
6/35	8/35	2/35	一日おいくらですか。	一泊おいくらですか。	5/35

町では、どちらに住むのがより好きですか。」は矛盾する文だと判断できる。このような危険性については、次のように考察をすることができる。

考察 2：コーパスの構造化のためには、コーパス中で欠けている文が、データスパースネスの問題により存在しないのか、文法・意味的に不可能な文なのを判断するため必要な情報の表示が望まれる。

8a. ～ 8b. では、コーパス中に多く現われる数字置換が観察される。

姉が二人います。 : 人数は二人です。
姉が三人います。 : 人数は三人です。

荷物は四つです。 : 荷物は四個です。
荷物は全て二つです。 : 荷物は全て二個です。

12a. ～ 14b. と 16a. ～ 16b. は「禁煙席・窓側の席・通路側の席」、19a. ～ 19b. は「クレジットカード・トラベラーズチェック」の入れ換えであり、意味的に類似な単語のクラスに分類可能である。他の例を挙げると、「めがね店・レコード店・土産店等」の様な単語クラスも本コーパス全体の中から類推関係により抽出できる。

3.3 構文論：肯定・否定・疑問文

7a. ～ 7c. と 10a. ～ 10b. の文では肯定・否定が対立している。また、9a. ～ 9b. では、以上に示した現在・過去の対立と、否定・疑問の対立が重なっている。コーパスの構造化を検討しつつ、この様な重なりを言語学理論に基づき考えると、ある文の分析とは、その文が参加できるすべての類推関係に対し、それぞれがどのような関係を表しているかを解明することだと考えられる。従って、次のように考察することができる。

考察 3：コーパスに欠けている文の成立可能性を評価するためには、その文が参加するすべての類推関係の観点から評価する必要がある。

3.4 意味：換言

12a. ～ 14b. と 16a. ～ 16b. では、依頼表現が二つの言い方で表現されている。「～にして下さい・～をお願いします。」の文である。本コーパスは依頼表現が豊富であるので次の様な言い方も存在する。

ライターをください。 : ライターがほしいのですが。
コショウをください。 : コショウがほしいのですが。
ビデオテープをください。 : ビデオテープがほしいのですが。
・・・

まけてください。 : まけてよ。
どいてください。 : どいてよ。

18a. ～ 18b. は意図的で等しい換言である。勿論、そのような対立文を特徴付けるためには人の判断が必要であるが、換言研究の第一段階として、類推関係抽出を用いることができると考えられる。

4 今後の課題

コーパスの構造化を目指して、抽出された多数の類推関係を一覧して検討できる表示形式が望まれる。付録の 3a. ～ 4b. の例に含まれる「いくら」という単語を含む文を、コーパス全体から短い順に数文抽出し、その類推関係を上記の 3 つの考察を考慮した表 3の形式で表示した。表 3のような行列の作成方法はまだ検討中であるが、この様に行列化することで、考察 1 に対応することが可能となる。

また、それぞれの行列のセルにコーパス中に実際に存在する文との間で成立する類推関係の数を追記した。その値は、考察 2 に対応して、欠けている文の成立可能性と相関があると考えることができる。例えば、図中の (a) の欠けている文「それはいくら。」の方が、(b) の文「あれでいくらですか。」より一般的表現であるが、類推関係の数も (a) の方が多い。

しかし、様々な文に対してその仮説が正しいことを示すためには、考察 3 に基づき、この「いくら」の観点の行列以外のすべての観点の行列でも、この傾向にあるかを検証する必要がある。

5 おわりに

本研究では、自動的コーパス構造化を目指して、類推関係抽出手法を提案した。「旅行会話基本表現集」のコーパスの一部分を使って、得られた対立の例を挙げた。また、類推関係行列の必要性を示し、そのような行列の表示ができれば、それぞれの文に起こる言語現象を明示することができることを示した。また、コーパスにある類推関係に基づいて欠ける文の追加可能性を示した。この類推という現象は普遍的であるから、本研究は、他の言語にもそのまま適用可能である。

謝辞

本研究は通信・放送機構の研究委託により実施したものである。

参考文献

Lloyd ALLISON & Trevor I. DIX
A bit string longest common subsequence algorithm
Information Processing Letter, 1986, Vol.23, pp 305-310.

Yves LEPAGE
Analogy and formal languages
Proceedings of FG/MOL 2001, Helsinki, 2001.

Yves LEPAGE & 白井 諭
類推に基づく構文解析の変定数の影響評価
言語処理学会第8回年次大会, けいはんなプラザ, 2002年3月, pp. 391–394.

Robert A. WAGNER and Michael J. FISCHER
The String-to-String Correction Problem
Journal for the Association of Computing Machinery, Vol. 21, No. 1, January 1974, pp. 168-173.

付録

1a. いつ。：なぜ。
1b. いつですか。：なぜですか。

2a. いつ。：どちら。
2b. いつですか。：どちらですか。

3a. いつ。：いくら。
3b. いつですか。：いくらですか。

4a. どちら。：いくら。
4b. どちらですか。：いくらですか。

5a. 気分が悪い。：気分が悪いのですが。
5b. 預けた荷物をもらいたい。：預けた荷物をもらいたいのですが。

6a. 大きいです。：小さいです。
6b. もっと大きいのがありますか。：もっと小さいのがありますか。

7a. 大丈夫です。：大丈夫ですか。
7b. 入院が必要です。：入院が必要ですか。
7c. お腹が一杯になりました。：お腹が一杯になりましたか。

8a. 十一時です。：十二時です。
8b. 一人部屋にしたい。：二人部屋にしたい。

9a. わかりません。：わかりました。
9b. いいえ、いきません。：いいえ、いきました。

10a. 入院が必要です。：入院が必要ですか。
10b. お腹が一杯になりました。：お腹が一杯になりましたか。

11a. 正装が必要ですか。：入院が必要ですか。
11b. 正装しなければいけませんか。：入院しなければいけませんか。

12a. 喫煙席にして下さい。：禁煙席にして下さい。
12b. 喫煙席をお願いします。：禁煙席をお願いします。

13a. 喫煙席にして下さい。：窓側の席にして下さい。
13b. 喫煙席をお願いします。：窓側の席をお願いします。

14a. 禁煙席にして下さい。：窓側の席にして下さい。
14b. 禁煙席をお願いします。：窓側の席をお願いします。

15a. 船便でお願いします。：航空便でお願いします。
15b. この手紙を船便でお願いします。：この手紙を航空便でお願いします。
15c. このはがきを船便でお願いします。：このはがきを航空便でお願いします。

16a. 窓側の席にして下さい。：通路側の席にして下さい。
16b. 窓側の席をお願いします。：通路側の席をお願いします。

17a. 英語の新聞はありますか。：日本語の新聞はありますか。
17b. 英語を話せる人に代わって下さい。：日本語を話せる人に代わって下さい。

18a. 英語を話せる人はいますか。：日本語を話せる人はいますか。
18b. 英語を話せる人に代わって下さい。：日本語を話せる人に代わって下さい。

19a. クレジットカードでいいですか。：クレジットカードをなくしました。
19b. トラベラーズチェックでいいですか。：トラベラーズチェックをなくしました。