A Study of the Number of Analogies contained in Large Multilingual Corpora

大規模多言語コーパスにおける類推関係の推算

イヴ・ルパージュ (Yves LEPAGE)
yves.lepage@atr.jp
国際電気通信基礎技術研究所 (ATR)

1 はじめに

言語学では、古くから「類推」が新たな文を解析や生成する方法の一つとして知られている[1]。古典ギリシャやローマ時代では、「アノマリア」[2]という概念で表層的には類推関係にならないタイプの意味的類推関係を示した。又、逆に、意味的には類推関係が満たされないが、表層的に類推関係が成立する場合があるため (例： *Abby is baking vegan pies.* : *Abby is baking.* :: *Abby is too tasteful to pour gravy on vegan pies.* : ??*Abby is too tasteful to pour gravy on.*) 、生成文法派の言語学者は、生成文法理論の立場から説明している。

本研究では、大規模コーパスにおける類推関係数を推定を試みた。「真の類推関係」のみを数える目的で、表層では類推関係が満たされたもののうち、同時に意味的にも類推関係になる場合を数える。表層レベルでは、提案済みの定義を採用する。意味のレベルでは、翻訳しても意味は文が異なっても伝わるという仮説を立て、ある言語での表層と意味の両方の類推関係が、他の言語へ翻訳する際にも、表層上の類推関係を満たす文として翻訳できるかを議論する。

2 コーパス

本研究では、「旅行基本表現集」（BTEC）を対象に分析を行なう。これは、旅行と観光に関する会話表現の広範囲でカバーする多言語コーパスである。音声翻訳に関する研究コンソーシアム C-STAR により[3]、162,318 文の表現が様々な言語に翻訳されている。本研究では、そのうち、中国語、英語と日本語のコーパスを使用する。中国語では 96,234 異なる文があるのに対し、英語では 97,769、日本語では 103,274 の文がある[4]。本コーパスの傾向を表 1の通りに示す。BTEC の文は平均的に短いと言える。

[1]主な参考文献は、(Paul 20, chap.5)、(Bloomfield 33, p.276)、(Mounin 68, p.119–120)、(Itkonen 94, p.48–50)、(Pullum 99, p.340–343)。

[2]ギリシア語の形容詞 $\alpha\nu o\mu\acute{\alpha}\lambda o\varsigma$ から。正しい分析は $\alpha\nu + \eth\mu\acute{\alpha}\lambda o\varsigma$（平かない、不似合い）。Lat. similis ↔ gr. $\eth\mu\acute{\alpha}\lambda o\varsigma$。そこから、英語の「anomaly」。

[3]http://www.c-star.org/.

[4]日本語の文の少しでも多い数は、漢字と平仮名での異なる書き方で説明できる（例：「ください」又は「下さい」。図 5にその実際例を示す。

3 表層的類推関係

表層的類推関係を次のように形式化できる[5]。

$$A : B :: C : D \quad \Rightarrow \quad \begin{cases} \mathrm{d}(A,B) = \mathrm{d}(C,D) \\ \mathrm{d}(A,C) = \mathrm{d}(B,D) \\ |A|_a + |D|_a = |B|_a + |C|_a, \forall a \end{cases}$$

ここでは、a は記号（文章で、文字）を示し、$A \sim D$ は記号列（コーパスで、文）である。記号 a が記号列 A 中に出現する頻度を $|A|_a$ で表す。また、$\mathrm{d}(A,B)$ は A と B の間で削除を 1・挿入を 1 のコストで[6]計算した編集距離である。この式が成立する場合なり $A \sim D$ が、表層的類推関係にあると言える。

4 類推関係文の抽出

4.1 表層的類推関係

それぞれの言語（中国語、英語、日本語）のコーパスを対象に表層的類推関係にある文数抽出した。表層的類推関係の例を図 3で示す。また、表 1はそれぞれの計算の結果である。これによると抽出された類推関係は大変多い。例えば、英語では約 10 万の文を含む約 250 万の類推関係が抽出できた。

図 1の左側は、コーパスに対する類推関係数の変化を示す。

一つの文あたりいくつの類推関係が存在しているか調査した結果、英語に対し、図 2の左グラフの通りとなった。右側は Zipf の法則に従った形になっているが、類推関係が多い文が類推関係数 750 と 2,000 の周りに不自然なピークがある。文の長さとその文に類推関係の数の検討をしても相関関係を見つかることができなかった。

4.2 意味的類推関係

自動的に意味的類推関係文を抽出するためには、意味の表示が必要なるが、検討対象コーパスには、そのような意味の表示はない。しかし、これは多言語コーパスであるので、いろいろな言語で対応している翻訳文の集合からなっている。翻訳

[5](Lepage 01)、FG/MOL。

[6]ここは、置換や変換は数えない。

文対の中に意味の表示がかくされていると考えられる。

単言語においては、表層的類推関係のうち、どれが意味的類推関係も満たすが判定することは困難である。しかし、ある言語で表層的にも、意味的にも、類推関係にある文はそれを他の言語に翻訳した場合でも表層的類推関係を満たすという仮説をたて、その妥当性を議論する。

$$\begin{aligned} \forall \mathcal{L},\ & \exists A \in \mathcal{L}(\text{`A'}), \\ & \exists B \in \mathcal{L}(\text{`B'}), \\ & \exists C \in \mathcal{L}(\text{`C'}), \\ & \exists D \in \mathcal{L}(\text{`D'}), \quad A : B :: C : D \end{aligned}$$

$$\Rightarrow \quad \text{`A'} : \text{`B'} :: \text{`C'} : \text{`D'}$$

すなわち、ある4つの文に対して、どの言語でも表層的類推関係を満たす翻訳が見つかると、意味的類推関係も満たす文であることが考えられる。

しかし、実際にこの仮説に基づき意味的類推関係を抽出する場合には、二つの問題がある。一つは、全ての言語でこのテストを行うことは困難である。もう一つは、ある言語でも、全ての可能な翻訳（換言）の作成は困難である。それに対して、本研究で使用するコーパスは換言文が少ない。英語の一つの文に対して中国語への可能な翻訳文は平均 1.20 文である。また、日本語への翻訳文は平均 1.52 文である。

この様な二つの問題はあるが、英語コーパスを対象に実験を行った。その表層的類推関係、2,384,202 関係の中に 238,135 関係は中国語でも満たされていることが分かった。その関係には 25,554 文が含まれている。従って、中国語訳の表層的類推関係も用いると、英語での表層的類推関係は約 10% は意味的類推関係であること、即ち「真の類推関係」であると考えられる。日本語訳の表層的類推関係を用いると、共通の類推関係は 336,287 関係（英語での表層的類推関係の 14%）であった。その関係には 13,602 文が含まれている。このような関係の例を図 4に示す。また、中国語訳にも日本語訳にも表層的類推関係にある場合は、13,602 文を含む 68,164「真の類推関係」が得られた。次章でこの低い価とその理由について議論する。

図 1の右グラフは抽出となる文数に対する「真の類推関係」の数の変化を表す。指数関数的な表層的類推関係の数の増加とは違って、線形的な成長となることが分かる。

5 考察

英語での表層的類推関係の数は、250 万あったが、このうち、日本語に表層的類推関係が成立するものは 30 万関係になり、又中国語でも成立するものは 7 万関係になった。ここからもれた類推関係は本当に意味的類推関係ないと言えるのであろうか。実際はその中に意味的類推関係が含まれていると考えられる。この原因としては、一文あたりの翻訳候補数が少ないことが考えられる。

手法を逆に考えると、即ち、二つの違う言語間で同時におこる表層的類推関係が意味的類推関係であるという仮定は類推に基づく翻訳の原理である[7]。この手法では、すくなくとも一つの表層的類推関係を持つ英文を日本語へ翻訳すると、平均 174 翻訳候補（換言文）が得る（正しくない文も含まれている）。実際にコーパスにある 1.52 文の候補と比較するのは難しい。この例を図 5で示す。

また、換言文を増やすと、類推関係の可能な組み合わせが増加する。全ての対象した言語で表層的類推関係が偶然に見つかる可能性も大きくなる。それに対して、対象する言語の数を増やさなければならないと考えられる。

6 おわりに

本論文では、実験により大規模多言語コーパスにおける類推関係の数を推定した。また、異なり 10 万文のコーパスで、「真の類推関係」、即ち表層でも意味でもの類推関係を推定した。その結果、約 1 万 4 千文を含む約 7 万の「真の類推関係」が得られることができた。本実験で用いたコーパスは、一文当たりに翻訳候補が少ないので、得られた数は非常に低い価になったと考えられる。

7 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

Leonard Bloomfield, *Language*, Holt, New York, 1933.

Esa Itkonen, Iconicity, analogy, and universal grammar
Journal of Pragmatics, 1994, vol. 22, pp. 37-53.

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

Yves Lepage, 比例類推的記号列形式言語, 言語処理学会第 7 回年次大会, 東京大学, 2001 年 3 月, pp. 93–96.

Georges Mounin, *Clefs pour la linguistique*, bibliothèques 10/18, Seghers, Paris, 1968.

Hermann Paul, *Prinzipien der Sprachgeschichte*, Niemayer, Tübingen, 5th ed. 1920 [1st ed. 1880].

Geoffrey K. Pullum, Generative grammar, *The MIT Encyclopedy of Cognitive Sciences*, The MIT Press, Cambridge, 1999, p. 340–343.

[7] (Lepage 01)、言語処理学会第 7 回年次大会。

	類推関係の数	類推関係を含む表現の数	コーパス中の表現の数	文の長さ（文字数）		
				平均	±	標準偏差
中国語	**3,567,123**	77,039	96,234	11.00	±	5.77
英語	**2,384,202**	53,250	97,769	35.14	±	18.81
日本語	**1,910,062**	53,579	103,274	16.21	±	7.84
「真の」類推関係	**68,164**	13,602				

表 1: 多言語コーパスにおける類推関係の数。

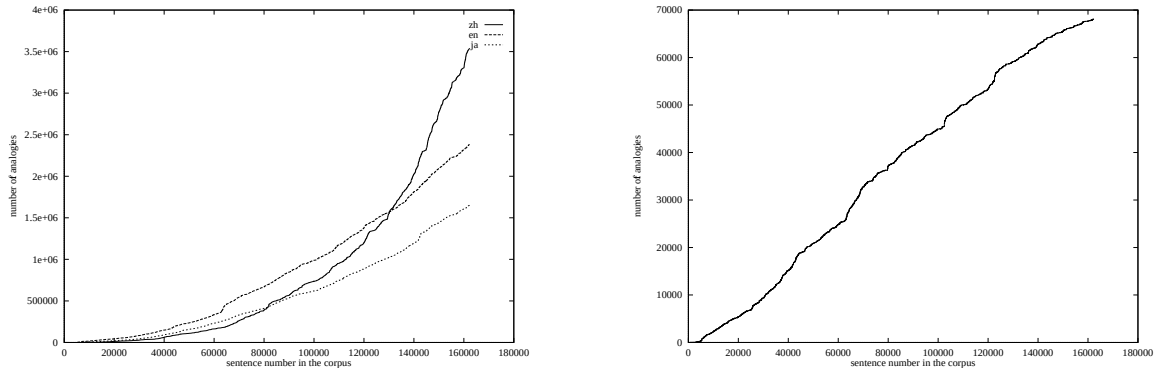

図 1: コーパス数に従い、類推関係数の変化。左：中国語と英語と日本語の表層的類推関係のみ。右：その言語で同時表層的類推関係（「真の類推関係」）数の変化。

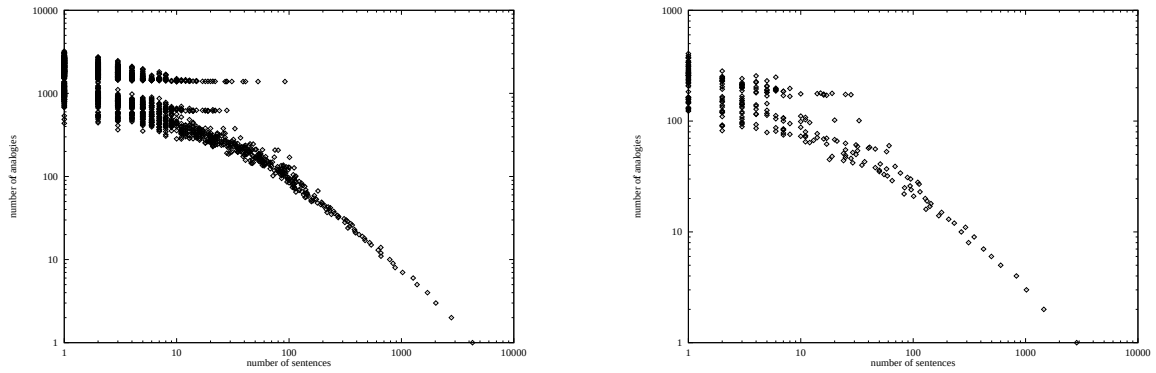

図 2: 1 文にあたり類推関係数の分布。左：英語の表層的類推関係のみ。右：中国語と英語と日本語の全てで層的類推関係（「真の類推関係」）数。横軸；表現の数（log スケール）。縦軸；類推関係の数（log スケール）。

美しい。	:	美味しい。	::	何にしますか。	:	何味にしますか。
お釣りをください。	:	お釣りをもらえますか。	::	住所を教えてください。	:	住所を教えてもらえますか。
保険をかけますか。	:	保険をかけたいのですが。	::	ヒルトンホテルに行きますか。	:	ヒルトンホテルに行きたいのですが。

図 3: 日本語での表層的類推関係の例。

メキシコ料理のほうが好きです。	:	中華料理のほうが好きです。	::	この辺りにメキシコ料理店はありますか。	:	この辺りに中華料理店はありますか。
↕		↕		↕		↕
I prefer Mexican food.	:	*I prefer Chinese food.*	::	*Is there a Mexican restaurant around here?*	:	*Is there a Chinese restaurant around here?*

図 4: 意味的類推関係になる 2ヶ国語での表層的類推関係の例。

コーパス中	
2×	日本の新聞をください。
1×	日本の新聞を見せてください。
1×	日本の新聞を見せて下さい。

類推に基づく翻訳結果	
210×	日本語の新聞をください。
177×	日本の新聞をお願いします。
147×	日本語の新聞をお願いします。
132×	日本の新聞をください。
105×	日本語の新聞をいただきたいのですが。
...	...
3×	日本新聞はありますか。
3×	日本新聞をいただけませんか。
3×	日本新聞サイズをお願いします。
3×	日本新聞何かをお願いします。
3×	日本分の新聞をお願いします。

図 5: 左側は英文 *A Japanese newspaper, please.* のコーパスにある実際日本語翻訳。右側は同文の頻度の最高と頻度の最低の 5 文の日本語への類推に基づく翻訳結果。合わせて、異なる翻訳候補 261 文が得た。