E-51

パラレルコーパスからの語彙的パラフレーズ獲得
Extracting Lexical Paraphrases from a Parallel Corpus

下畑 光夫[1]
Mitsuo Shimohata
mitsuo.shimohata@atr.co.jp

隅田 英一郎[1]
Eiichiro Sumita
eiichiro.sumita@atr.co.jp

1 はじめに

ある事柄、意図を表現する場合に、自然言語では数多くの表現で表すことができる。このような同じ内容を持つ表現群をパラフレーズと呼ぶ。パラフレーズは、コミュニケーション支援や自然言語の機械処理のために有用であり、近年その研究が盛んになってきている[1]。

本論文では、パラレルコーパスから語彙的パラフレーズを獲得する方法について述べる。 パラレルコーパスとは、ある言語の文(原文)と他言語の対訳文から構成されるコーパスを指す。等しい訳文を持つ複数の原文は基本的に同じ意味を持つ(同義文)とすると、これらの原文集合は同義文集合を形成する。同義文の対から、DPマッチングにより局所的な差異を抽出することで語彙的なパラフレーズを獲得することができる。本手法は深い言語知識を使用しないため、品詞タグ付きコーパスのみから獲得することができ、構文情報や記載されている語の対訳情報などは必要としないという特長がある。

また、獲得したパラフレーズを用いて文を書き換えた場合に、同義性が保持されるかどうかを評価した実験について報告する。日英のパラレルコーパスから、日本語、英語の双方について同一アルゴリズムでパラフレーズを獲得し、それぞれの言語について評価を行った。

2 対象となるパラフレーズの特徴

本論文が獲得対象としているパラフレーズは、以下の2点の特徴を持っている。

語彙的パラフレーズ

パラフレーズは、その差異が及ぶ範囲により構文的パラフレーズと語彙的パラフレーズの2つに大別することができる。構文的パラフレーズとは、差異が広範囲(例えば複数文節)にわたるものであり、能動・受動の違いなどが該当する。語彙的パラフレーズとは、差異が局所的(例えば文節内)であり、同義語の入れ換えなどが該当する。本論文では、その差異が2語以内であるパラフレーズが語彙的パラフレーズであるとし、獲得対象としている。

文脈情報の包含

パラフレーズの同義性は文脈に大きく影響される。同義文間の差異だけを抽出してパラフレーズとすると、文脈に関する情報が全く欠落してしまう。そこで、獲得するパラフレーズには、差異だけでなくその前後の1語も文脈的情報として取り入れる。例えば、“は(ありますありません)か”は前後の語を含んだパラフレーズの一例である。“あります”と“ありません”という2語は、語単独ではほとんどの場合は同義ではないが、前後の語を制約することで適切なパラフレーズとなっている。

3 パラフレーズ獲得手法

パラフレーズは、以下の手順により獲得する。

同義文集合

パラレルコーパスから同一の訳文を持つことを条件に同義文集合を形成する。日本語パラフレーズの場合、英訳文が一致する日本語文を集めて集合を形成することに相当する。

同義文対からの表現対抽出

各同義文集合から、すべての組合わせの同義文対を取り出す。それらの同義文対について語を単位としてDPマッチングを適用し、共通する語ならびに異なる語を判定する。図1の上部に英語の同義文対にDPマッチングを適用した例を示す。2文間で対応する語は線で結ばれている。2文間で異なる語が多い(本手法では2語以上と定義)同義文対は構文的パラフレーズを含む可能性が高いため、処理対象から除外する。

残った同義文対から、異なる語とそれらの前後の共通する語を取り出し、表現対を抽出する。図1の下部に、上部の同義文対の差異から抽出した表現対を示す。

フィルタリング

前段で抽出された表現対を、2種類の統計的判定基準により選別する。一つは表現対の頻度である。出現した同義文集合が3未満の表現対は除外する。もう一つは、表現対の頻度と表現対を構成する2表現の頻度の比率である。表現対の頻度がどちらの表現の頻度の5%に満たない場合は除外する。この2つ

[1]ATR音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Labolatories

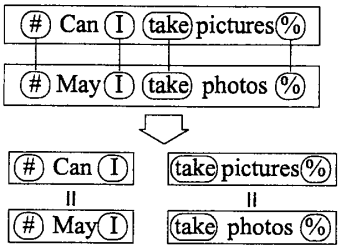

図 1: 同義文対からの表現対の抽出

いい	です	か
いい	のです	か
いい	でしょう	か
いい	のでしょう	か
#	トイレ	は
#	お手洗い	は
#	化粧室	は
#	Could	you
#	Would	you
#	Can	you
#	Will	you
a	guarantee	%
a	warranty	%

図 2: 獲得したパラフレーズ

の選別条件を共に満たした表現対がパラフレーズ対として認定される。

さらに、パラフレーズ対は推移関係を利用してパラフレーズ集合にまとめる。例えば、表現 1 = 表現 2 かつ 表現 2 = 表現 3 ならば、表現 1, 表現 2, 表現 3 は一つのパラフレーズ集合を形成する。

4 獲得したパラフレーズの同義性評価

旅行会話を対象とした日英のパラレルコーパス [2] を用いて実験を行った。3 語以下からなる短い文を除いた日本語 102,406 文、英語 97,092 文を学習データとし、同一のパラメータで日本語、英語双方のパラフレーズを獲得した。その一部を図 2に示す。図中、“#” は文頭を、“%” は文末を表している。また、各パラフレーズ集合の最上段はコーパス中で最頻の表現を表している。これを標準パラフレーズと定義する。

獲得されたパラフレーズを用いてテストデータ文の書き換えを行った。書き換えは、非標準表現を標準表現に置換することで行った。そして、書き換え前後の文を評価者に与え、書き換え前の文と基本的意味において同義であるかどうか、また書き換え文自体に文法的誤りが生じていないかを判定した。評価結果を表 1に示す。表中、“書き換え率” は全テスト文のうち書き換えられた文の割合を、“同義” は書き換えが正しく行われた率を、“非同義” は書き換えにより同義とならなかった率を、“文法エラー” は書き換え文に文法的誤りが生じた率を表す。

表 1: パラフレーズの同義性評価結果

	日本語	英語
テスト文数	8,092	7,564
書き換え率 (%)	13.0	17.1
同義 (%)	93.5	83.2
非同義 (%)	6.1	6.4
文法エラー (%)	0.4	10.4

この実験結果より、日本語では 93%、英語では 83% の精度で同義性を保持しながら書き換えることができた。英語は日本語と比較すると、文法的に誤りのある文が多く生成されている。例えば、主語の単複を同一視するパラフレーズや、“# Are you” = “# Do you” といったパラフレーズなどが適用されると、文法的誤りが発生することがある。

5 まとめと今後の課題

本論文では、パラレルコーパスから語彙的パラフレーズを獲得する方法について述べた。本手法は、品詞タグ付きコーパスに適用することができ、構文情報や対訳情報などの深い言語知識を必要としないという利点がある。また、日英のパラレルコーパスから日本語と英語それぞれについてパラフレーズを獲得し、その同義性が日本語で 93.5%、英語で 83.2% 保持していることを示した。

本手法で獲得されるパラフレーズの量や精度は、パラフレーズの対象となる言語と訳文の言語の関係に大きく影響されると考えられる。今後は、様々な 2 言語を組み合わせてパラフレーズの獲得ならびに評価実験を行い、言語の性質と得られるパラフレーズの関係について分析していきたい。

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

[1] 乾健太郎. 言語表現を言い換える技術. 第 8 回言語処理学会チュートリアル, pp. 1–21, 2002.

[2] T. Takezawa. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd LREC*, pp. 147–152, 2002.