

接頭・接尾辞を含む数量表現の翻訳方法

延原 由高 池原 悟 村上 仁一

鳥取大学工学部

{nobuhara,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

現在、数量表現の日英翻訳では英語解析の立場において助数詞とターゲット（対象とする名詞）の関係から翻訳パターンを決める方法が考えられ [1]、日本語解析の立場からは、構造的に分類する方法や、数量詞とターゲットの意味関係による翻訳パターンの研究が行われてきた [2]。

しかし、これらの研究は数詞、助数詞、名詞の組み合わせのみを対象としており、接頭・接尾辞は数量詞と名詞の組み合わせによる翻訳に直接影響をおよぼさないため翻訳の対象とされていなかった。

また、接頭・接尾辞は新聞記事などにおいて数多く出現し、訳は曖昧になるケースが多い。そこで、本研究では新聞記事10000文の対訳コーパスから、助数詞、名詞の意味属性、および数詞の桁数に着目した接頭・接尾辞の翻訳方法を提案し、その評価を報告する。

2 接頭・接尾辞の分類

2.1 接頭・接尾辞の定義

従来、数量詞に対する接頭・接尾辞は数量接頭語、数量限定語とされている [3] が、その明確な定義は決まっていない。そこで、本研究では日英機械翻訳システム ALT-J/E の日本語形態素解析部分をパッケージ化した ALTJAWS の品詞コードを接頭・接尾辞の定義に用いた [4]。ALTJAWS の品詞コードは接辞だけで20種以上あり、品詞は細かく分類されている。本研究では ALTJAWS における品詞コード 6240、6170 を接頭・接尾辞として扱う。以下に数量詞の定義と接頭・接尾辞の定義を示す。

数量詞 = [数量接頭語][符合](数)(助数詞)[数量限定語]

表1:接頭・接尾辞の定義

	品詞コード	名称	例
接頭辞	6240	助数詞継承型	約、平均など
接尾辞	6170	前置助数詞型	弱、以下など

2.2 接頭・接尾辞の出現頻度

接頭・接尾辞を含む数量詞を辞書や新聞記事から抽出した。その結果の頻度を表2に示す。表2より、接頭・接尾辞を含む文は新聞記事では4.7%、辞書においては0.2%であった。また、新聞記事において接頭・接尾辞を含む数量詞は全体の3割以上と頻度が高い。頻度が高い理由は接頭・接尾辞が新聞記事の経済欄に多く含まれるためである。そこで、本研究では新聞記事10000文の対訳コーパス（以下対訳コーパス）を調査した結果を利用して、接頭・接尾辞の翻訳方法を提案する。

表2:接頭・接尾辞の出現頻度

コーパスの種類	全体の文数	数量詞の頻度	接頭・接尾辞の頻度
辞書	25000文	832件(3.3%)	51件(0.2%)
新聞記事	10000文	1315件(13.2%)	471件(4.7%)

2.3 意味による接頭・接尾辞の分類

接頭・接尾辞の意味の違いから翻訳方法を考案するために、対訳コーパス(全471件)の接頭・接尾辞を「数の概略」、「数の上限」、「数の順序」、「年月の割合」、「数の性質」の5種類に分類した。以下に例を示す。

- 数の概略(4種):-約(122件)、前後(14件) など
- 数の上限(9種):-以上(33件)、以下(10件) など
- 数の増幅(6種):-増(32件)、強(5件)、弱(5件) など
- 年月の割合(9種):-月産(12件)、年間(10件) など
- 数の順序(1種):-第(7件)
- 数の性質(8種):-計(7件)、当り(1件) など

次章で上記の順(「数の性質」は除く)に従って翻訳方法を示す。

3 接頭・接尾辞の翻訳方法

対訳コーパスより抽出した37種類、471件の接頭・接尾辞のうち、本研究では15種類307件を翻訳対象とした。残り22種類の接頭・接尾辞は頻度が少ないため翻訳の対象外とする。以下に本研究で翻訳の対象外とした接頭・接尾辞の例を示す。

例：余り(2件)、当り(2件)、およそ(1件)など

3.1 「数の概略」の翻訳方法

3.1.1 「約」の翻訳方法

対訳コーパスにおいて接頭辞「約」の翻訳結果は主に「about」となる。しかし、英文において「about」の直前に前置詞「at」がくる場合、「about」は省略される[5]。これには以下の2つの理由がある。

- 1: 「at」、「about」と前置詞が2つ並ぶと冗長になる
- 2: 「at」、「about」は表す数の曖昧性が違う

以上の点を考慮して図1に「約」の翻訳方法を示す。

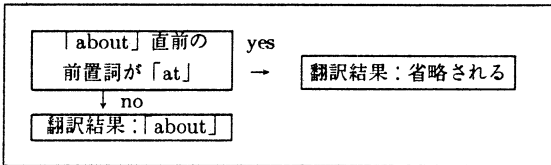


図1: 「約」の翻訳方法

3.1.2 「前後」の翻訳方法

対訳コーパスを調べた結果、接尾辞「前後」は主に「around」、「about」のどちらかに翻訳されることがわかった。「around」は「about」より数を表す範囲が曖昧[6]である。概略の数量詞は桁数の0が多ければ数が曖昧になる。そこで本研究では以下の式(1)を用い数の概略の基準値を示す。

$$\text{数の曖昧性の基準値} = \frac{\text{桁数} - 0 \text{の桁の個数}}{\text{桁数} + 1} \quad (1)$$

例: 123456000 → $\frac{6}{10} = 0.6$ (0に近づくほど曖昧)

例: 100000000 → $\frac{1}{10} = 0.1$

式(1)を対訳コーパスに適応し、基準値を定めた結果、数詞の桁数が0.5以上なら「前後」の訳を「about」とし、以下なら「around」となった。

以上の点を考慮して図2に「前後」の翻訳方法を示す。

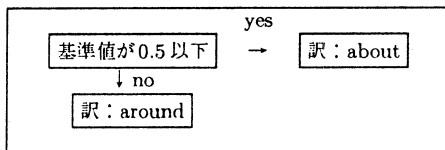


図2: 「前後」の翻訳方法

3.1.3 「程度」の翻訳方法

接尾辞「程度」には意味が(1)「根本にかかわることではなく、程度がどうということ」と(2)「程度の大小が物事の適否にかかわること」の2種類がある[7]。対訳コーパスを調べた結果、「程度」が(2)

の意味を持つとき訳は省略される可能性が高いことが示された。以下に「程度」の2種類の意味を示す。

意味(1)例:株が10億円程度暴落した。

意味(2)例:砂糖を10g程度ならあげる。

また、対訳コーパスから概略の意味を持つ接頭・接尾辞のうち数量の範囲を表す「数詞-数詞+助数詞」を表現することができるのは「程度」のみであった(例:20-30%程度)。そのときの翻訳結果は「数量詞-数量詞+range」となる。

以上の点を考慮して図3に「程度」の翻訳方法を示す。

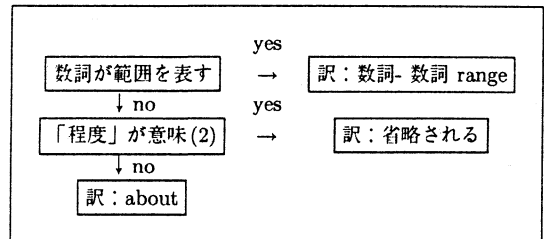


図3: 「程度」の翻訳方法

3.2 数の上限に関する翻訳

日本語で「10以上」と言うときは10が含まれるが、英語で以上を意味する「more than ten」、「over ten」には10は含まれない。厳密に「10以上」というときは、「ten or more」、「ten or over」といい、厳密にいう必要のない場合は「more than ten」、「over ten」を使うことが可能である[8]。

例: 彼女はケーキを10個以上食べた。

→ She ate ten cakes or more.

→ She ate more than nine cakes.

→ She ate over nine cakes.

上記の例より、「more than」、「over」をつかって厳密に「~以上」と表す場合、規定の数から1を引かなければならない。しかし、対訳コーパスにおいて「more than」、「over」は厳密に数を表す必要のない場合に使用されている。つまり、1程度の数の影響が表れにくい桁数の大きい場合に適用されていると考えられる。そこで、対訳コーパスから「more than」、「over」を適用する桁数を調べ、10以下を基準値として翻訳に適用する。この「more than」、「over」の基準値は「less than」、「under」にも適用できる。

また、「over」、「under」は年齢、「below」、「above」は気温の意味を表す助数詞に使われる。以上の点を考慮して表3、表4に「以上」、「以下」の翻訳方法を示す。

表3: 「以上」の翻訳方法

桁数10以下	助数詞が温度	助数詞が年齢	英訳
○	×	○	or over
○	○	×	above
○	×	×	or more
×	×	○	over
×	○	×	above
×	×	×	more than

表4: 「以下」の翻訳方法

桁数10以下	助数詞が温度	助数詞が年齢	英訳
○	×	○	or under
○	○	×	below
○	×	×	or less
×	×	○	under
×	○	×	below
×	×	×	less than

3.3 「数の増減」の翻訳方法

3.3.1 「強、弱」の翻訳方法

接尾辞「強、弱」はそれぞれ主に「more than」、「over」、「less than」、「under」の2種類に翻訳される。「over」、「under」は桁数が大きい場合に使用されることが多い[6]。対訳コーパスを調べた結果、翻訳の基準となる数詞を「一億以上」と定め、数詞が「一億以上」なら「強、弱」の訳は「over」、「under」、以下なら「more than」、「less than」とする。以下に例を示す。

例:100億円強 → over one billion yen

例:1000円強 → more than thousand yen

以上の点を考慮して表5に「強、弱」の翻訳方法を示す。

表5: 「強」、「弱」の翻訳方法

	桁数1億以上	翻訳結果
強	○	more than
	×	over
弱	○	less than
	×	under

3.3.2 「増」の翻訳方法

対訳コーパスを調べた結果、「増」の訳は「数詞+increase」となる。しかし、数量表現が「数量詞+の+名詞」(QのN型)、「名詞+数量詞」(QN型)の場合、「数量詞+over+名詞」の方が「~の量を増えている」という意味を持つため「over」を用いる。以下に例を示す。

例:80%増 → increase 80 %

例:前年度比80%増 → 80 % over the previous

以上の点を考慮して図4に「増」の翻訳方法を示す。

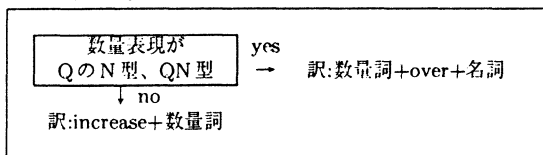


図4: 「増」の翻訳方法

3.4 「年月の割合」の翻訳方法

接頭辞「年産」、「月産」、「年間」、「月間」、「年」、「月」で訳として扱われるのはそれぞれ、「per year」、「per month」、「a month, a year」であるが「per」は商業的な意味合いを持つときに使用される[9]。対訳コーパスを調べた結果、助数詞が「ドル」、「円」などの金額の意味を持ち、かつ数量詞と共に起する名詞が商業的な意味合いを持つ場合、翻訳に「per」を用いることがわかった。本研究では商業的な意味を持つ名詞の定義に岩波書店より出版されている日本語語彙体系[10]の意味属性番号を適用する。なお、適用した意味属性番号は「1867(取り引き)」、「1920(労働)」以下に属する名詞である。意味属性番号とは約40万語の名詞を最大12段、2710の属性に分類した木構造の番号である。以上の点を考慮して表6に「年月の割合」の接尾辞の翻訳方法を示す。

表6: 「年月の割合」の翻訳方法

助数詞が金額	意味属性番号が「1867」、「1920」	翻訳結果
○	○	per month, per year
○	×	a month, a year
×	×	a month, a year

3.5 「数の順序」の翻訳方法

対訳コーパスでは、接頭辞「第」が来る場合、通常その数詞は序数詞(例1)となる。しかし、助数詞が「号」、「番」など番号を表す場合は「No+数詞」(例2)となり、共起する名詞の意味属性番号が「憲法(1163)」以下に属する場合は「Article+数詞+of+名詞」となる(例3)。以下に例を示す。

例1: 第二ラウンド → the second round

例2: 第30号 → No 30

例3: 憲法第9条 → Article 9 of the Constituion

以上の点を考慮して表7に「第」の翻訳方法を示す。

表7: 「第」の翻訳方法

助数詞が番号	名詞の意味属性番号(1163)	翻訳結果
○	×	No+数詞
×	○	Article+数詞+of+名詞
×	×	序数詞

4 評価実験

3章で作成した翻訳方法の精度を以下の2種類の対訳コーパスを使用して評価する。また、本研究で提案した翻訳方法の有効性を評価するために頻度が最も高い翻訳を正解とする翻訳方式(以下デフォルト値)とも比較する。

- 1: 新聞記事に含まれる数量表現(クローズドテスト)
- 2: 機能試験文集¹に含まれる数量表現(オープンテスト)

¹日英機械翻訳の機能をテストするために表現を集めた約6200文

なお、新聞記事に含まれる数量表現の評価実験は翻訳方法作成のデータと同一のためクローズドテストとなり、機能試験文集中に含まれる数量表現の評価実験は翻訳方法作成のデータと無関係のためオープンテストとなる。評価方法と評価を以下に示す。

$$\text{正解率} = \frac{\text{正しい翻訳結果}}{\text{全データ}} \quad (2)$$

表：8 評価実験

	本研究の翻訳方法	デフォルト値
-	正解率	正解率
新聞記事 頻度	58% 167/287	39% 112/287
機能試験文 頻度	53% 22/41	44% 18/41

4.1 評価結果

評価実験の結果を以下に示す。

- 新聞記事で正解率58%、デフォルト値の正解率39%を得られた。正解率はデフォルト値より18%上昇した。
- 機能試験文集中で正解率53%、デフォルト値の正解率44%を得られた。正解率はデフォルト値より9%上昇した。

上記の結果より、本研究で考案した翻訳方法はデフォルト値より正解率が高いことがわかり、有効であるといえる。

4.2 成功例と失敗例

今回の実験の成功例と失敗例を以下に示す。

1) 成功例

100 億円以上 → more than 10 billion yen

2) 失敗例 1: 翻訳結果が異なる場合

約三十メートル前方の敵兵

○ an enemy about 30 meters ahead

× an enemy *some* 30 meters ahead

3) 失敗例 2: 翻訳結果が省略される場合

約十七万五千トンの年間生産能力

○ an annual capacity of about 1750,000 tons

× an annual capacity of 1750,000 tons

5 考察

5.1 評価結果について

オープンテスト、クローズドテストを行った結果、本研究で提案した翻訳方法の正解率はそれぞれ、58%、53%であった。これより本研究の正解率はデフォルト値よりも10%程度上昇することがわかった。しかし、正解率は50%程度と低い。これは接頭・接尾辞の訳が一

意に定まる条件が少ないためと、一意に定める必要がないためだと考えられる(失敗例1参照)。しかし、本研究で提案した翻訳手法は接頭・接尾辞の翻訳の一般常識を取り入れているためより自然な訳が期待できる。

5.2 省略条件について

本研究の翻訳方法は主に接頭・接尾辞の訳し分けを扱ってきた。しかし、接頭・接尾辞の訳はそれ自身意味を持たないため省略されることが多い(失敗例2参照)。対訳コーパスにおいては全体の2割近くが省略された訳となった。訳が省略される条件のひとつとして、接頭・接尾辞の重複があげられる。以下に例文を示す。

例：約80%強 → more than 80 %

例：最大約10億円 → maximum of ten billion yen

接頭・接尾辞が重複し、どちらかが省略される場合どちらの訳が文章に必要とされるかを考慮する必要がある。

6 あとがき

本研究では接頭・接尾辞の翻訳方法を提案した。そのために、ALTJAWSの品詞コードから接頭・接尾辞を定義し、数詞の桁数、名詞の意味属性、助数詞に着目することで翻訳方法を提案した。

正解率はオープンテスト58%、クローズドテスト53%の結果が得られ、デフォルト値よりも10%程度上昇することが解った。また、より高い精度を求めるために、接頭・接尾辞が省略される条件や他の標本の翻訳方法を取り入れる必要がある。

参考文献

- [1] Bond: "When and How to Disambiguate?" PROCEEDING OF THEMIDDIM-96 SEMINARL (1996)
- [2] 旭岡他: "日本語数量表現の意味処理について"、自然言語処理研究会 Vol90.No78.P129-136(1990)
- [3] 吉武、高山、鈴木、太細: "日本語数量表現とその英語への機械翻訳に関する一考察"
- [4] 宮崎、白井、池原: "言語過程説に基づく日本語品詞の体系化とその効用"、自然言語処理論文誌, Vol.2, No.1, pp.12-21(1995.1)
- [5] 小西 友七: "英語の前置詞" (1976)
- [6] 横山、落合: "曖昧な数量詞を含む名詞句の解析法" (1995)
- [7] 新村 出: "広辞苑 第五版", 岩波書店 (1998)
- [8] 富井 篤: "英語数量表現辞典" (1995)
- [9] 長谷川、アラン、橋本: "プロシード和英辞典" (1988)
- [10] 池原、宮崎、白井、横尾、中岩、小倉、大山、林(1997)"日本語語彙体系"、岩波書店