2005－NL－170（26）
2005／11／22

対訳例文から中日数量表現の翻訳規則の獲得について

邵敏　　尹大鵬　　任福継　　黒岩眞吾
知能情報工学科　徳島大学

中日機械翻訳における数量詞の処理は誤りが多い．本研究では,それらの文法特徴に基づき量詞を分類して処理する方法を提案する．まず,中日対訳コーパスから収集した数量詞の例文を形態素解析して,得られた量詞の種類と数量詞に修飾される名詞の語義特徴を統計する,そして異なる数量詞と出現する位置の異なりなどにより,機械翻訳における数量詞の翻訳規則を獲得する．これに基づいて構築した翻訳実験システムは 2 つのモジュールによって構成される,一つはこの数量詞を翻訳するかどうかを確認する．もう一つは,数量詞を翻訳する場合,翻訳形式を選定する．得られた翻訳規則を利用して中日数量詞の機械翻訳の評価実験を行った．最後に,実験データの適応性を検証し,提案した方法の有効性を検証した.

Rule-based translation of quantifiers for Chinese-Japanese machine translation

Min Shao　Dapeng Yin　Fuji Ren　Shingo Kuroiwa
Department of Information Science and Intelligent Systems
University of Tokushima

Quantifiers and numerals often give rise to trouble in Chinese-Japanese machine translation. In this paper, an approach is proposed based on the syntactic features after classification. First, morphological analysis is performed on the sentences extracted from a Chinese-Japanese aligned corpus, which consists of quantifiers and numerals. Next, statistical information is obtained based on the word meaning of the noun that has an accompanying quantifier. Using the difference in quantifier type and position between Chinese and Japanese, quantifier translation rules were acquired. The translation and experiment system is made up of 2 modules. One is to check the quantifier translation and the other is to select the correct translation rule. The evaluation experiment was conducted using the acquired translation rules. Finally, the adaptability of the experimental data is verified and the validity of the proposed method is proven.

1. はじめに

近年,中日機械翻訳研究が盛んに行われている.しかし中日両言語の数量表現についてはあまり研究されておらず,適切な翻訳が行えない場合が多い,そこで本稿では量詞の中日翻訳における検討を行う.中日両言語の量詞を比較して,全体として中国語の量詞は日本語より多いことが分かる.一般的に,中日両言語の数量詞表現は以下の三種類のパターンに分類することができる.

1．　中国語文には数量表現があり、対応する日本語文にも数量表現がある.
2．　中国語文には数量表現があり、対応する日本語文には数量表現がない.
3．　中国語文には数量表現がなく、対応する日本語文には数量表現がある.

本稿では,これらの三種類と量詞の位置情報及び名詞との関係を利用して,機械翻訳における中日数量詞の翻訳規則をまとめ,評価実験を行った.

2. 翻訳システム

2.1. 数量詞の分類

数量詞の種類には以下の三つのパターンがある.

1. 「一」＋量詞
2. 「一」以外の常数＋量詞
3. 概数＋量詞

数量詞の出現頻度は表１に示す.

表１ 数量詞の出現頻度

数量詞	「一」＋量詞	「一」以外の数詞＋量詞	概数＋量詞
個数	3685	491	807
百分比	74%	10%	16%

また,一つの数量詞がセンテンスにおける出現位置によって表現する意味がかなり違う.文における出現位置によって,それらを以下の種類に区分する.

1. 文頭 （ 数量詞＋ …）
2. 文中 （… ＋数量詞＋ …）
3. 文末 （… ＋数量詞 ）

2.2. 提案する翻訳モデル

我々は量詞の分類規則を利用し,「一」＋量詞と「一」以外の常数＋量詞と「概数」＋量詞どれに属するかを確認する,それから量詞の種類を確認して,次は数量詞出現する位置情報を判定する.システムの流れを図１に示す.

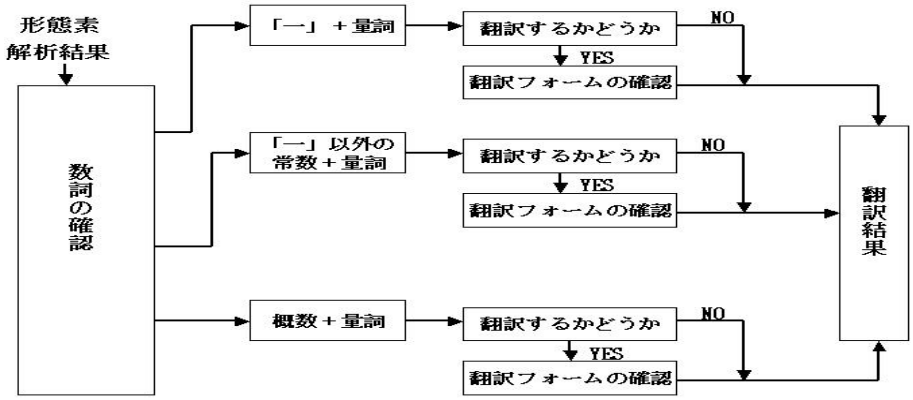

図１中日数量詞機械翻訳システム流れ

3. 数量詞の翻訳ルール

3.1. 「一」＋量詞の翻訳ルール

中国語の量詞には様々な種類がある.名量詞（集合量詞、度量衡量詞、時間量詞、重ねる量詞）,動量詞などである.現代中国語の特徴によって「一」＋量詞というパターンをA.「一」＋名量詞, B.「一」＋ 動量詞の二種類に分類した.

表２ 数量詞における量詞の比率

数量詞	名量詞	動量詞
個数	3036	643
百分比	83%	17%

数量詞を分類したコーパスにおける「一」＋名量詞と「一」＋動量詞の示す比例を表２に示す.このデータはそ

れぞれの翻訳ルールと大きく関わる.本節では,それを利用してそれぞれの翻訳ルールを説明する.

3.1.1 名量詞の翻訳ルール

1)「一」＋　量詞が文頭に現れる場合

◆　翻訳する

➢　「一 」+ 量詞+（修飾語）＋名詞　——中国語

「一 」+ 量詞＋の＋名詞＋格助詞　——日本語

例：一颗枪弹落在街心.

一発の銃弾が道の真中に落ちていた.

2)「一」＋　量詞が文中に現れる場合

◆　翻訳しない

➢　代名詞/前置詞/方位詞/場所詞/副詞/名詞+（助詞）+「一」+ 量詞+（修飾語）+名詞　——中国語

例：我被领到楼梯下面一间昏暗的房子里.

階段の下の暗い部屋へ案内された.

◆　翻訳する

➢　動詞 +（代名詞/助詞/前置詞）+ 「一」+ 量詞+（修飾語）+ 名詞+接続詞+ 「一」+ 量詞+（修飾語）+ 名詞　——中国語

「一」+ 量詞＋の＋（修飾語）＋名詞＋と＋「一」+ 量詞＋の＋（修飾語）＋名詞　——日本語

例：桌子上放着一对大烛台和一个大香炉.

テーブルに一対の大きな燭台と一つの大きな香炉がおいてある.

3)「一」＋　量詞が文末に現れる場合

◆　翻訳する/翻訳しない

➢　主語に当たる名詞が一つではない時＋(連用修飾語)+(動詞)+ 「一」+ 量詞+（間投詞）——中国語

名詞１+助詞+名詞２+助詞＋…数量詞+ずつ　——日本語

例：　蟹三明治和鲜橘子汁各一个啊.

クラブサンドとオレンジジュース一つずつですね.

3.1.2　動量詞の翻訳ルール

1)「一」＋　量詞が文中に現れる場合

◆　翻訳しない

➢　代名詞+ 「一」+ 量詞+ （修飾語）+名詞　——中国語

例：他开始明白了人生是怎么一回事,做一个人究竟应该怎样.

彼にはやっと人生とはいかなるものか、人はいかに生くべきかなどが分りはじめたようだ.

2)「一」＋　量詞が文末に現れる場合

◆　翻訳しない

➢　動詞+（助詞）+「一」+ 量詞+（間投詞）　——中国語

例：总之,他并不动一下.

相変らずその場所に立って動かない.

3.2　「一」以外の数詞＋量詞の翻訳ルール

「一」以外の数詞とは具体的な数字をさす. これらの数量詞はほとんど翻訳する.

◆　翻訳する

1)「一」以外の常数＋量詞が文頭に現れる場合

➢　数詞+量詞+（修飾語）+名詞　——中国語

数詞+量詞+の＋（修飾語）＋名詞　——日本語

例：　两件新衣服.

二着の新しい服.

2)「一」以外の常数＋量詞が文中に現れる場合

➢　動詞/代名詞/前置詞/形容詞/方位詞/接続詞/名詞+（助詞）+数詞+量詞+（修飾語）＋名詞　——中国語

数詞+量詞+の＋（修飾語）＋名詞　——日本語

例：打坏了两株松树.

二株の松の木が倒れていた.

3)「一」以外の常数＋量詞が文末に現れる場合

➢　動詞/前置詞フレーズ/副詞+（量詞）+数詞+量詞+（間投詞）　——中国語

数詞+量詞+助詞＋動詞　——日本語

例：死了将近五年.

死んでからまだ五年とたたない.

◆　翻訳しない

➢　「一」+「两」+量詞+名詞

例：他有时候掉过头来问她一两句话.

彼は時々ふり返って何かいった.

3.3 概数＋量詞の翻訳ルール

概数とは話し手が正確な数を知らないか, 或いは言いたくないか, または言う必要のない場合におよその数を言うときに使われる表現である. 概数＋量詞が中国語の量詞の中で 16%ぐらい占めるが, 概数＋量詞の構成は複雑である. 概数を表す語は次の二種類に分類できる. 第一は疑問詞「几」と「许多」,「很多」,「无数」,「若干」＋量詞を構成した. 第二は不定数量詞, 不定の数量を表す.「些」「点儿」の二つがある.「些」「点儿」の前に数詞は「一」しか使えず,「一些」、(いくらかの)「一点儿」(少しの) のようになる.

4. 日本語翻訳形式の選定

我々は 3 節を通じて, 中日機械翻訳における中国語の文にある数量詞を翻訳するべきかどうかについて言及した. 次は該当する数量詞がどのような形式で訳すかを確認する. 中国語の常用数量詞は約 150 である. それに対して, 日本語の数量詞は約 50 である. 中国語の数量詞を日本語に翻訳するとき, 必ずいくつかの中国語の数量詞が一つの日本語の数量詞と対応することになる.

また, 修飾される名詞の違いによって, 一つの中国語の数量詞がいくつかの日本語の数量詞と対応することもあり得る. 例えば, 中国語では, 名詞「毛巾」(タオル)、「船」(船) を修飾する数量詞は「条」であるが, 日本語では, タオルには「枚」、船には「隻」が使われる.

上述のような一対一、一対複数、複数対一の中日数量詞の対応関係を解決するために, 大量のコーパスにおける数量詞のデータを統計し. その結果数量詞は修飾される名詞の特徴に制限されることが分かった. 次は名詞の特徴

を分析し,中日機械翻訳における150中国語の数量詞の対訳ルールのサンプルを表3に示す.一行目は中国語量詞,二行目は対応する日本語の量詞の種類の数である.三行目は中国語の名詞の集合N1である.四行目はN1と対応する日本語量詞の翻訳形式Ｔ１（Q+量詞、Qは数量詞の数詞を表わす）である.五行目は中国語の名詞集合N2である.六行目はN2と対応する日本語量詞の翻訳形式T2（Q+量詞）である.

表3 中日量詞対訳表

中国語量詞	翻訳種類	中国語名詞集合N1	日本語量詞T1	中国語名詞集合N2	日本語量詞T2	…	中国語名詞集合Ｎｎ	日本語量詞Ｔｎ
条	4	毛巾	Q+枚	皮带,蛇	Q+本	…	船	Ｑ＋隻
只	12	鸟类	Q+羽	船	Q+隻	…	小動物	Ｑ＋匹
件	2	衣服	Q+着	小事	Q+つ			
..	..	..	..	..	…	..	…	…

まず,表 3 の中日量詞対訳表の一行目から対応する中国語量詞を確認する.それから「中国語名詞集合Ｎｎ」から修飾される中国語名詞をマッチングする.もし見つかれば,対応する日本語翻訳形式 Tn を確認し,翻訳結果を出力する.

実際,中国語の名詞の数は膨大になるため,全部対訳表に登録することができない.この問題を解決するために“語義類似度”を使用する.中日量詞対訳表の中国語名詞集合Nn中に日常的,典型的な名詞を用例とする.「知网」の語義辞書と意味素辞書によって修飾される名詞と中日数量詞対訳表にある用例を語義類似度で計算する.類似度が高ければ高いほど同じ量詞を使うことになる.

我々は一つの具体的な例を通して,語義類似度計算方法を詳しく説明する.

例：　　树上有一只百灵鸟.

形態素解析した後,数量詞「一只」と修飾される名詞「百灵鸟」を得る.次に中日量詞対訳表に量詞「只」をマッチングする.その後,中国語名詞集合N1-Nn中のあらゆる中国語名詞との類似度を計算する.

例えば,表3中日量詞対訳表の「只」と対応する中国語名詞集合N1は「鸟类」であり,「鸟类」をキーワードして,中国語語義辞書の中で,マッチングする.結果としては以下に示す.

鸟类---- N-----bird|禽

語義情報は「bird|禽」である.次にこの語義情報はキーワードになって,中国語意味辞書中に探して,探した結果は以下に示す.

823---- bird|禽 ----820

この記録の葉ノードは823,父ノードは820である.

先の探した方法を利用して,表3中日量詞の訳す表の「只」の記録中国語名詞集合N1-Nn中に名詞すべて計算して,葉ノードと父ノードは確認し,まだ例1の修飾された名詞「百灵鸟」はキーワードになって,中国語語義辞書中に探して,語義情報は得る,「禽」ではる,この語義情報はキーワードになって,中国語意味辞書中に探して,葉ノードと父ノードは確認し.もし表3中日量詞対訳表の「只」の記録中国語名詞集合N1-Nn中にすべて名詞の葉ノードと父ノードは例中に「百灵鸟」の葉ノードと父ノードと同じの場合は,この名詞対して日本語翻訳形式Tを利用して.まだ同じのノードがなしの場合,これらのノードと「百灵鸟」のノードの間に絶対的な距離を計算して,最短い絶対的な距離のノードの名詞の翻訳形式は利用する.

具体的な語義類似度計算公式は： $d = Dis\ (W_1, W_2)$ …(1)

公式(1)中 W1 と W2 はそれぞれ 2 つの意味素のノードの位置.

$$Sim\ (p_1, p_2) = \frac{a}{d + a} \cdots (2)$$

公式(2)中 2 つの意味素（primitive）を表して,d は P1 と P2 の意味素の段階の体系中のパスの長さで,この数が整数です. α は 1 つの調節することができるパラメーターで,私達が実験を通して, α が 1.6 がなのであることを確認する.

$$Sim\ (W_1, W_2) = \max_{i=1\ldots n,\, j=1\ldots m} Sim\ (S_{1i}, S_{2j}) \cdots (3)$$

2 つの中国語の語句 W1,W2 について,もし W1 は n 語義解釈の項目（概念）があるならば：S11,S12,……S1n, W2 は m 語義解釈の項目（概念）がある：S21,S22,……S2n,私達は定めて,W1 と W2 の相似度(度合)それぞれの語義類似度が極みは大いに値する.すべての概念がすべて最後に帰結してもとは義で表すため,だから意味素の相似度(度合)が計算するのが概念の相似度(度合)計算の基礎だ.すべての意味素が上によって位から下りて関係して 1 つの木形の意味素の段階の体系を構成するため,私達のここは語義類似度計算する方法まで簡単な語義を通じて(通って)を採用する.d だまで 2 つの意味素のこの段階の体系の中のパスを仮定して,公式 (3) によって,私達はこの義原の間の語義の距離を得ることができる.

5. 実験と評価

本節では上述した翻訳ルールに基づいて数量詞の機械翻訳評価実験を行う.　実験に使用した中日対訳コーパスは数量詞を含む 5000 文である.これらの対訳コーパス「坊ちゃん」「鼻」「斜陽」「家」の 4 冊からなる.その内の 3000 文を用いて,開発した翻訳システムを利用して実験を行った（評価基準は数量詞）.結果を表 4 に示す.

表 4　評価実験結果

文	成功	失敗
3000	2619	381
	（87.3%）	（12.7%）

その内 2619 件が翻訳成功し,381 件が失敗した.失敗の主な原因としては形態素解析のミスによるものであった.

6. まとめ

本稿では我々が提案した分類方法によって,数量詞を「一+量詞」、「一以外の数詞+量詞」、「概数+量詞」に分類した.また,修飾語と被修飾語との関係によって,量詞は名量詞と動量詞に分けられ,また文における出現位置によって,翻訳するかしないかを確認した.そして翻訳するとき,我々が統計した中日数量詞対訳表によって,具体的な翻訳形式を確定した.

最後に,本研究において開発した評価システムと実験結果を示した.実験結果より,提案手法を用いた中日機械翻訳における数量詞の処理の有効性を示すことが出来た.

参考文献

[1]朱学峰，王惠：现代汉语量词与名词的子类划分，现代汉语语法信息词典技术报告之二

[2]飯田朝子：日本語主要助数詞の意味と用法,東京大学大学院人文社会系研究科博士論文ライブラリー (1999)

[3]Qun LIU , Sujian LI　“Word Similarity Computing based on HowNet,” Computational Linguistics and Chinese Language Processing　Vol. 7, No. 2 , August 2002, pp. 59-76