社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180 (28)
2007／7／25

Super-Functionを用いた日英機械翻訳における日付・時間表現の抽出

篠山 学† 土屋 誠司†† 黒岩 眞吾†† 任 福継††

† 徳島大学大学院工学研究科
†† 徳島大学大学院ソシオテクノサイエンス研究部
770-8506 徳島県徳島市南常三島 2-1
E-mail: †{sasayama,tsuchiya,kuroiwa,ren}@is.tokushima-u.ac.jp

あらまし 用例に基づく機械翻訳の一つであるSuper-Function（SF）に基づく機械翻訳は，名詞を変数化することで用例の適用範囲を広げられるという特長を持つが，名詞以外の表現を含む日付・時間表現では，日付・時間表現全体をひとつの名詞として抽出することができず，その数字部分しか変数化できないという問題があった．この問題を解決するため，本稿では，日付・時間表現を抽出する手法を提案する．SFに基づく機械翻訳では名詞を抽出するために名詞判定規則を用いている．また抽出した各名詞の言語間の対応を得るために単語辞書を用いている．本手法ではまず名詞判定規則に日付・時間表現を抽出する規則を追加し日付・時間表現を抽出した．次に抽出した日付・時間表現を日英に共通な形に変換することで日付・時間表現の対応を得た．作成した規則を用いて評価実験を行ったところ日本文で適合率96.7%，再現率98.2%，英文で適合率94.7%，再現率92.7%を得られた．
キーワード 日付・時間表現，Super-Function，機械翻訳

Extracting Date/Time Expressions in Super-Function based Japanese-English Machine Translation

Manabu SASAYAMA†, Seiji TSUCHIYA††, Shingo KUROIWA††, and Fuji REN††

† Faculty of Engineering, The University of Tokushima
†† Institute of Technology and Science
2-1 Minamijosanjima, Tokushima 770-8506
E-mail: †{sasayama,tsuchiya,kuroiwa,ren}@is.tokushima-u.ac.jp

Abstract Super-Function Based Machine Translation(SFBMT) which is a type of Example-Based Machine Translation has a feature which makes it possible to expand the coverage of examples by changing nouns into variables, however, there were problems extracting entire date/time expressions containing parts-of-speech other than nouns, because only nouns/numbers were changed into variables. We describe a method for extracting date/time expressions for SFBMT. SFBMT uses noun determination rules to extract nouns and a bilingual dictionary to obtain correspondence of the extracted nouns between the source and the target languages. In this method, we add a rule to extract date/time expressions and then extract date/time expressions from a Japanese-English bilingual corpus. The evaluation results shows that the precision of this method for Japanese sentences is 96.7%, with a recall of 98.2% and the precision for English sentences is 94.7%, with a recall of 92.7%.
Key words Date/Time Expression, Super-Function, Machine Translation

1. はじめに

対訳コーパスを用いた機械翻訳には統計に基づく機械翻訳手法[9]と用例に基づく機械翻訳手法[1],[2],[7],[8]の二つに大きく分類される．対訳コーパス自体を翻訳に用いる利点は，自然文をそのまま翻訳規則としているため流暢な訳文を生成できることにある．しかし自然文を直接利用しているため，用例と入力文のわずかな表記の違いによって正しく翻訳できない場合がある．そこで用例に基づく機械翻訳のひとつであるSFに基づく機械翻訳[3]では，対訳コーパスをSFを用いて関数化して翻訳に利用することで適用範囲を広げ汎用性を高めている．機械翻訳におけるSFは原言語と目標言語の対応を示す関数とし

て定義されている．SF を用いることにより，ほとんどの機械翻訳システムに必須である構文解析を必要とせず処理の簡略化が可能となる．SF に基づく機械翻訳に類似した手法としては，テンプレートを用いた翻訳手法[10],[11] がある．これらは構文解析を使用しており，構文解析を使用せず対訳コーパスから直接 SF を作成している点において，SF を用いた翻訳手法とは異なる．他に類似の手法として，文献[1] では入力文の名詞すべてを考慮しているのに対し，SF を用いた翻訳手法では原言語と目標言語の両方に出現した名詞だけを扱っているため適切な訳文を得られると考えている．

SF を用いた翻訳手法には，対訳コーパスをそのまま関数化しているため主語や目的語の欠落に対して頑健という特長がある．このため省略の多い日常会話の翻訳に適している．しかし日常会話に頻出する要素である日付・時間表現については，数字や名詞部分のみしか抽出できないため副詞や前置詞など含んだ表現を変数化することができなかった．

用例ベース機械翻訳において日付・時間表現は文献[5] や文献[6] で行われているように数字や曜日，月などを変数化する場合，日付・時刻を示すタグを付与している．しかし前置詞等を考慮していないため日付と時刻が別々に扱われている．そのため多数のパターンが必要になる．

本論文では，SF に基づく機械翻訳のために日英対訳コーパスから日付・時間表現全体を抽出する方法を提案する．具体的には日付・時間表現の中に現れる副詞や前置詞なども名詞として扱い，日付・時間表現全体を一つの名詞として抽出する．そのために，日英ともに日付・時間表現のための規則を追加する．

抽出した日付・時間表現は日英に共通な表現（共通形）に変換しておく．これらの処理により名詞以外の品詞を含む日付・時間表現全体をひとつの名詞として扱い変数化することができる．

本手法は対訳コーパスから SF を自動抽出する処理[15] に組み込んだ．この処理は以下の二つの処理から成っている．1）各対訳文を名詞とそれ以外の部分に分ける．2） 単語辞書を用いて日本語名詞と英語名詞を対応付けし，日本語名詞の英文中における位置情報を得る．1）に日付・時間表現の規則を追加し，2）の単語辞書を用いる代わりに共通形への変換処理を行う．以下，2 章でこれまでの研究である SF に基づく機械翻訳や SF の抽出方法の概要を説明する．3 章では SF に基づく機械翻訳における日付・時間表現の問題点を述べた後に，提案手法を説明する．4 章で評価実験を行い提案手法の有効性を考察する．

2. Super-Function に基づく機械翻訳

SF とは原言語と目標言語の対応を示す関数である．SF では文は定数と変数から成ると定義する．変数を置き換えることで複数の異なる文を得ることができる．本論文では変数を名詞とし，その他の部分を定数とする．原言語と目標言語の変数と定数の関係を示すために SF の例を用いて説明する．

SF : X_{j1} は X_{j2} まで X_{j3} に乗った．

$\Longrightarrow$ X_{e1} took X_{e3} to X_{e2} .

J : [彼] は [駅] まで [タクシー] に乗った．

表 1 ノードテーブル (左) とエッジテーブル (右)
Table 1 NTB(left) and ETB(right)

J	E	L_j	L_e	Condition
ϕ	ϕ	1	1	1p
は	took	2	3	a
まで	to	3	2	the
に乗った.	.	-	-	-

E : [He] **took** [a taxi] **to** [the station].

この例では名詞を [ ] で囲ってある．英語の冠詞は名詞の一部として扱う．太字部分（名詞以外）がノードを表す．

SF はノードテーブル（NTB）とエッジテーブル（ETB）で構成される．表 1 はノードテーブル（左）とエッジテーブル（右）の例を示したものである．ノードテーブルは日本文と英文から抽出された各ノードの対応を表す．エッジテーブルは名詞の条件と日本語名詞の英文における位置情報（L_j, L_e）を保持する．条件の '1p' とは代名詞のコードを表す．

2.1 翻訳処理

SF を用いた日英機械翻訳の流れを次に示す．

（1） 日本文を形態素解析し，エッジ（名詞）とノード（その他の文構造）に分割する．

（2） ノードをノードテーブルと照合し，マッチした SF から対応する目標言語のノードを得る．

（3） エッジをエッジテーブルと照合し，各名詞の目標言語における順序を得る．

（4） 名詞を日英単語辞書で翻訳する．

（5） ノードとエッジを並びかえる．

2.2 コーパスから Super-Function の自動抽出

対訳コーパスから SF を自動抽出する処理の概要を図 1 に示す．図中の点線部分の処理に提案手法を追加する．

最初に，準備した対訳コーパス内の対訳文の日本文，英文を各々形態素解析し，名詞（名詞を修飾する形容詞を含む）とその他の部分に分ける（図 1 (1)）．形態素解析には日本文は「茶筌」[12]，英文は「Brill's tagger」[13] を用いる．

次に名詞に対し対訳単語辞書を用いて日本語名詞と英語名詞とを対応付けし，日本語名詞の英文中における位置情報を得る (図 1 (2))．この位置決定の処理を以下名詞照合と呼ぶ．(詳しくは 2.2.2 節で説明する)

名詞照合では対応する名詞がない場合も存在する．この対応する名詞がない名詞を孤立名詞と呼び，孤立名詞処理を行う (図 1 (3))．(詳しくは 2.2.3 節で説明する)

以上の処理を行うことでノードテーブル，エッジテーブルが作成できSF として抽出される．最後に抽出した SF は共通する日本語ノード毎にまとめる．尚，実際の SF の構成は図 2 に示す形式で表される．図中の四角の点線で囲われた部分が SF である．(1) と (2) がノードテーブルを表している．(3-a) と (3-b) がエッジテーブルで，(3-a) が名詞の位置情報，(3-b) が名詞の条件を表す．各条件は英文中の名詞の出現順に「:」で区切られて並んでいる．この例では「英文中の 1 番目の名詞は日

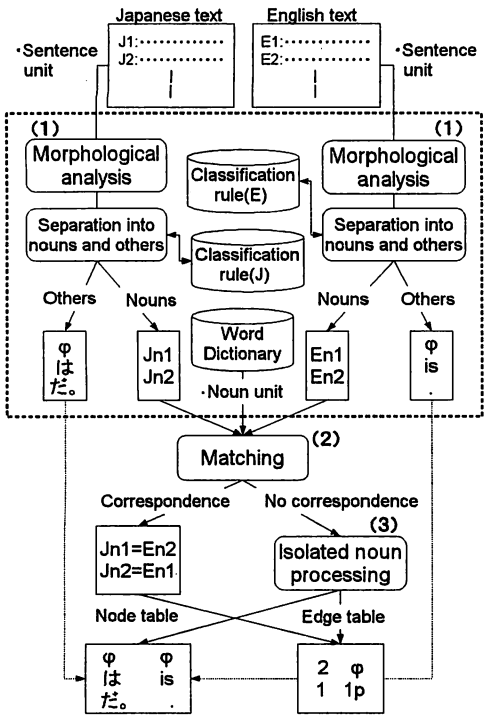

図 1 SF の抽出処理概要

Fig. 1 Outline of extraction process of SF

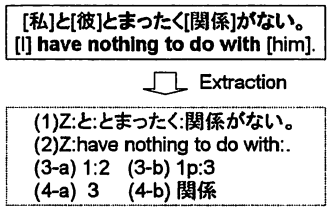

図 2 SF の構造

Fig. 2 Structure of SF

本文中の 1 番目の名詞で，その名詞の条件は '1p' である．英文中の 2 番目の名詞は日本文中の 2 番目の名詞で，その名詞の条件は '3'（目的格）である.」ということが分かる．(4-a) と (4-b) は孤立名詞とその位置情報を表す．この例では，孤立名詞が日本文中の 3 番目の名詞「関係」であることを表している．

2.2.1 形態素の分類

タグ付けされた形態素を名詞とその他の部分に分ける処理を説明する．この処理では品詞タグとそれを分類するために作成した名詞分類規則を用いる．この分類規則は日本文と英文で異なる．名詞分類規則にはある一つの形態素が名詞か否かを判断するための規則を記述している（表 2）．同規則では表 2 に示すように判断する形態素の以降（または以前）の形態素の品詞によって名詞か否かを判断している．例えば，サ変接続の次の形態素が助詞であった場合はそのサ変接続の形態素は名詞と判断する．現在，名詞分類規則としては 32 の規則を用いている．すべての形態素を名詞（エッジ）とその他の部分（ノード）に

表 2 名詞分類規則の一部 (日)

Table 2 A part of classification rule(J)

入力形態素	次の形態素	判定結果	例
名詞	助詞	名詞	駅 に
サ変	動詞	動詞	勉強 する
サ変	動詞以外	名詞	勉強 が
サ変	名詞	名詞	勉強 期間

分け終わったら，名詞は文中での出現順に並べておく．その他の部分は孤立名詞処理を経てノードテーブルを構成する．

2.2.2 名詞照合

上記の方法で抽出した日本文中の名詞と英文中の名詞に対して名詞照合を行い，日本語名詞の英文中における位置情報を得る．対応する英語名詞がなかった場合，日本語名詞が孤立名詞となり，対応する日本語名詞がなかった場合，英語名詞が孤立名詞となる（2.2.3 節）．

2.2.3 孤立名詞処理

孤立名詞処理について説明する．孤立名詞処理は孤立名詞が日本語名詞か英語名詞かによって処理が異なる．まず孤立名詞が英語名詞である場合，その英語名詞はエッジではなくノードとみなしノードへの編入処理を行う．例を次に示す．

元の英語：[We] **has a lot of** [snow] **last year.**

[We] の編入処理後：**We has a lot of** [snow] **last year.**

このようにエッジをノードに編入することで英語の形式名詞なども扱うことができる．

次に孤立名詞が日本語名詞である場合，その日本語名詞は孤立名詞として位置情報と共に SF 中の名詞の条件として登録する．これは日本語が名詞であっても英文中では名詞以外である場合が存在するためである．図 2 が孤立名詞を含む SF の例である．この例では「関係」は名詞（エッジ）と判定される．しかし「関係」に対応する英語名詞がないため,「関係」が孤立名詞となっている．

3. 日英 SF における日付・時間表現の抽出

本論文で提案する日英 SF における日付・時間表現の抽出処理は SF の自動抽出処理の内部に追加する（図 1 の点線部分）．日本文の処理は名詞とその他の部分に分ける処理に追加する．英文の処理は名詞とその他の部分に分ける処理へ追加すると共に形態素解析済みの文に日付・時間表現用の品詞タグを追加する．これは英語の形態素解析における詳細な情報が日本語の形態素結果に比べ不十分なためである．この処理概要を図 3 に示す．図中の点線で囲われた部分が日付・時間表現の処理部分である．囲われていない部分は既存の処理である．あらかじめ日本文，英文ともに各名詞分類規則に日付・時間表現抽出用の名詞分類規則を英語表現の本[16] から作成し追加しておく．まず名詞分類規則から日付・時間表現を名詞として抽出する．次に抽出した日付・時間表現を日英に共通な形式（以下，共通形と呼ぶ）に変換する．これらの処理により日付・時間表現を名詞として抽出し変数化することができる．以下，各処理について

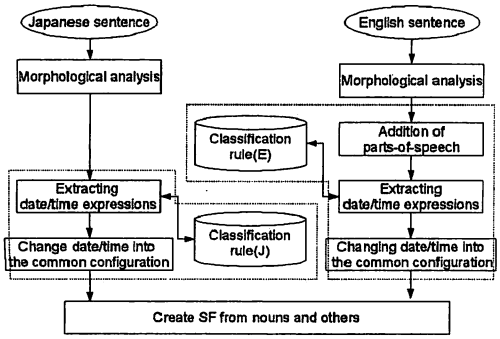

図 3　日付・時間表現の抽出処理概要

Fig. 3　Outline of extracting date/time expressions

日本文の場合と英文の場合に分けて説明する．

3.1　日付・時間表現の定義

本研究では，日付・時間表現を次のように定義した．

- 日本語日付表現

絶対的な表現（年月日と曜日）を日付表現とする．和暦の年号も含める．範囲表現は個別の部分を日付表現とする．相対的な表現（前日，今晩など）や曖昧な期間を表す表現（数年後，上旬など）は扱わない．

- 英語日付表現

日本語と同様に年月日と曜日を日付表現とする．月日には主に 5 つの表記（5th of May，5 May，5th May，May 5th，May 5）がある．前置詞「on」,「in」まで含めて日付とする．範囲表現，相対的な表現や曖昧な期間を表す表現は扱わない．

- 日本語時間表現

X 時 Y 分，X 時，X 時半の 3 種類を時刻とする．X と Y はそれぞれ任意のアラビア数字または漢数字を表す．これに接尾する品詞として，茶筌の「名詞-接尾-副詞可能」タグが付く可能性のある語のうち「前」,「過ぎ」,「ごろ」の 3 語を用いる．時刻の前に付く語として「午前」,「午後」の 2 語を用いる．

- 英語時間表現

日本語時間表現に対応した箇所を時間表現とする．例えば「X o'clock」や「half past X」などである．AM，PM も時刻の範囲とする．

3.2　共 通 形

共通形とは日英で異なる表記で表される日付・時間表現を同じ表記に変換した形を表す．共通形は年，月，日，曜日，時刻の順に「1776,7,4,Thursday,10:30」のように表される．曜日がない場合や時刻がない場合などは「1776,7,4,,」のようにそのまま出力する．

英文で日付・時間表現に含める前置詞は「in」,「on」,「at」の 3 種類と「before」,「about」,「around」の合計 7 種類とした．「in」は年や月のとき用いられ，「on」は特定の日に「at」は時刻に対して用いられる．また「before」,「about」,「around」は日本語の「前」,「ごろ」などに対応している．そのためこれらの前置詞を日付・時間表現として変数化しても，翻訳の時に場合わけして前置詞をつけることができる．

表 3　追加した名詞分類規則の一部 (日)

Table 3　Added noun classification rules(J)

入力形態素	次の形態素	例
名詞-数	名詞-接尾-助数詞	2　時
名詞-副詞可能	名詞-数	7月　3
名詞-接尾-助数詞	名詞-数	時　2
名詞-副詞可能	助数詞	時　前
名詞-接尾-副詞可能	助数詞	分　過ぎ

表 4　DATE タグの付与規則の一部

Table 4　Additional rule of DATE tag

D1	on/IN */NNE
D2	*/NNE
D3	on/IN the/DT */CD */CD of/IN */NNM
D4	on/IN */CD */CD of/IN */NNM
D5	on/IN the/DT */CD of/IN */NNM

表 5　TIME タグの付与規則の一部

Table 5　Additional rule of TIME tag

T1	at/IN */CD
T2	half/* past/JJ */CD
T3	*/CD past/JJ */CD
T4	*/CD after/IN */CD
T5	(*/DT)quarter/* to/TO */CD

3.3　日本文の日付・時間表現処理

表 3 に追加した分類規則の一部を示す．この分類規則に一致した形態素は日付・時間表現であると判定する．例えば形態素が「2」，品詞が「名詞-数」の場合，次の形態素の品詞が「名詞-接尾-助数詞」であれば形態素「2」は日付・時間表現と判定する．ただし「名詞-接尾-助数詞」の形態素が「時」,「時半」,「年」,「日」,「分」の場合であるときに限る．日付・時間表現が連続して出現する場合は一つの日付・時間表現として扱う．(例えば「1776年7月」や「8時30分」など) また日付・時間表現の間に「の」が出現する場合は「の」も日付・時間表現として扱う．(例えば「7月4日の木曜日」や「1999年の4月」など) 最後に抽出した日付・時間表現を共通形に変換する．この処理は日付と時刻それぞれ別に行う．変換後ひとつにまとめて共通形を完成させる．なお，すべての漢数字は全角数字に変換しておく．

3.4　英文からの日付・時間表現処理

英文では日本文と同様の日付・時間表現処理を行う前に品詞タグの追加を行う．英文は形態素解析によって品詞タグが付与されている．この品詞タグに日付・時間表現のためのタグ（日付表現は 'DATE'，時間表現は 'TIME'）を追加する．

この 'DATE' タグの追加規則の一部を表 4 に示す．また 'TIME' タグの追加規則の一部を表 5 に示す．表中のタグ 'NNE' は曜日を表し，'NNM' は月を表す．'CD' は数字（アルファベットを含む）のタグである．*は任意の数字，アルファベットを表す．() は出現しない場合を含めることを示す．この追加規則

表 6 追加した名詞分類規則の一部 (英)
Table 6 Added noun classification rules(E)

入力形態素	次の形態素	判定結果
NN	DATE	CUT
NN	TIME	CUT
NNS	DATE	CUT
NNS	TIME	CUT
NNP	DATE	CUT

に一致した場合，すべてのタグを 'DATE' もしくは 'TIME' に変更する．例えば「on/IN May/NNM 5/CD」のとき 'D7' の規則に一致するため「on/DATE May/DATE 5/DATE」とタグが 'DATE' に変換される．'TIME' も同様に変換する．タグを追加した結果,「,」の前後が 'DATE' なら「,」を'DATE' にする．

'TIME' タグ追加規則において，任意の数字が「-」で区切られている場合がある．(例えば「six-five」) このとき「-」の右側の数字が序数であれば分数を表すことになるため 'TIME' タグを付与しない．表 5 の規則のほかに*/CD の前後に名詞がない場合 'TIME' タグを付与する．これらの処理を前処理として行った後，日付・時間表現処理を行う．

表 6 に追加した英文用の分類規則の一部を示す．日本文の分類規則と同じくこの分類規則に一致した形態素は日付・時間表現であると判定する．また日本文の処理と同様に日付・時間表現が連続して出現する場合は一つの日付・時間表現として扱う．(例えば「on July 4」や「at 8:30」など) なお，アルファベットはすべて数字に変えておく．

最後に抽出した日付・時間表現を共通形に変換する．

4. 評価実験と考察

日付・時間表現がひとつの名詞として抽出され共通形に変換されれば日付・時間表現に対応した SF は作成できる．そこで提案手法により日英の日付・時間表現を過不足なく抽出し共通形に変換できるか否かを評価する．

4.1 実験条件

テストセットを対訳コーパスからランダムに選択したのでは十分な量の日付・時間表現が含まれない．そこで最初に正規表現を用いて日付・時間表現を含む可能性のある文を収集した．これらの文の中には日付・時間表現ではない文が含まれている．日付・時間表現を含む可能性のある文とは日本文において，「時」,「年」,「月」または「(日月火水木金土) 曜」が出現する文を表す．20 万対訳コーパスから日付・時間表現を含む可能性のある文を探した結果，4176 対訳文を得ることができた．4176 文中 100 文を規則の修正に用いた．テストセットには残りの 4076 文からランダムに選んだ 500 文を用いた．正解データ (共通形) は人手で作成した．日本文 500 文中に含まれる日付・時間表現数は 385 箇所であった．英文 500 文中に含まれる日付・時間表現数は 384 箇所であった．日本文，英文それぞれから適合率と再現率を求めた．適合率と再現率は次の式で表される．

表 7 実験結果
Table 7 Experimental result

	precision(%)	recall(%)
Date/time expressions(J)	96.7(378/391)	98.2(378/385)
Date/time expressions(E)	94.7(356/376)	92.7(356/384)

$$\text{適合率} = \frac{\text{正解した日付・時間表現数}}{\text{提案手法が抽出した日付・時間表現数}} \quad (1)$$

$$\text{再現率} = \frac{\text{正解した日付・時間表現数}}{\text{すべての日付・時間表現数}} \quad (2)$$

一文に必ずしも日付・時間表現が含まれていない場合や複数個含まれている場合もある．一箇所の数え方は隣接していれば一箇所とし，そうでなれば別々に数える．

4.2 実験結果

表 7 に結果を示す．日本文では適合率 96.7%，再現率 98.2%となり日付・時間表現を高精度で認識可能であることが分かった．英文では適合率 94.7%，再現率 92.7%となり日本文には劣っているものの高い精度で日付・時間表現を得られていることが分かった．日本文での正解数は 378 箇所，失敗数は 13 箇所であった．13 箇所のうち 11 箇所はルール不足によるもので，2 箇所は誤抽出だった．英文での抽出正解数は 384 箇所，失敗数は 20 箇所であった．20 箇所のうち 15 箇所はルール不足によるもので，5 箇所は誤抽出だった．

抽出に成功した日本文，英文の例を示す．入力文，日付・時間表現，共通形の順に表示する．

例)「彼女は１９９０年７月１７日の午前６時に生まれた.」
→「１９９０年７月１７日の午前６時」
→「1990,7,17,,6:00」
例)「She/PRP was/VBD born/VBN at/IN six/CD AM/NNP on/IN July/NNM 17/CD ,/, 1990/CD ./.」
→「at/TIME six/TIME AM/TIME on/DATE July/DATE 17/DATE ,/DATE 1990/DATE」
→「1990,7,17,,6:00」

この例から SF を作成した場合，次のようになる．形式は図 2 と同様である．ϕは該当無しを示す．名詞の条件の 'T' は日付・時間表現を示す．

(1) Z:は:に生まれた.
(2) Z:was born:.
(3-a) 1:2 (3-b) Z:T
(4-a) ϕ (4-b) ϕ

これにより，例えば「彼は十月に生まれた.」という文もこの SF に一致し，翻訳することができる．

4.3 考察

抽出に失敗した箇所について，失敗例を挙げながら日本文と英文に分けて検証する．

- 日本文

失敗例 1 :「一時的」の「一時」を時間と判定した．
形態素解析結果が「名詞-数 (一) ＋名詞-接尾-助数詞 (時) ＋名詞-接尾-形容動詞語幹 (的)」となったため「名詞-数＋名詞-接尾-助数詞」が名詞分類規則と一致し誤抽出された．

失敗例 2 :「二人前」を日付と判断した.
日付・時間表現以外の語句が「名詞-数+名詞-接尾-助数詞+名詞-副詞可能」の規則に適用されてしまったため誤りとなった.

失敗例 1 は慣用句の一部に用いられている時間表現を誤抽出した例である. 日付・時間表現に接尾辞が付属する場合の規則を追加すれば解決できると考えられる. 失敗例 2 では規則適用後に表層語を用いて日付・時間表現かどうかを判断する必要がある.

- 英文

失敗例 3 :「How about 12:45」の「about」を時間と判定した. 「about */CD」の規則に適用されてしまうため「about」が時間と判断された.

失敗例 4 :「Monday , Wednesday and Friday」のうち「Monday , Wednesday」が一つの日付と判断された.

失敗例 3 の場合の解決方法としては「How about」を別扱いする規則を追加することが考えられる. また失敗例 4 の場合もカンマを用いた並列の表現と認識できていないため, 並列を認識するための規則を追加する必要がある.

失敗例 5 :「One Tuesday」の one を日付と判定した.
この例は「ある火曜日」の「one」が日付と判断された. 英文では「one」は不定代名詞や一般の人, 前述の名詞の代用など日付・時間表現以外での用法が多い. また単に数字を表す場合と時間を表す場合の区別は文脈から判断しなければならない場合がある. この問題に対応するために SF を抽出する際, 日付・時間表現として扱い, 該当する共通形が日本文になかった場合, 通常の名詞として処理するようにすることを考えている.

日付・時間表現に含めなかった前置詞として「by」と「until」がある. 含めなかった理由は日本語の表現「まで」ひとつに対して「by」と「until」はそれぞれ終了までの期限と継続した時間を区別して表すからである. そのため日本文から形態素だけを用いて「by」と「until」の区別をつけることは困難であると考えられる.

日本文において日付・時間表現に含めなかった表現で頻出していたのは「来週の, 来月の, 来年の, 次の」などである. これらは英語の「next」と対応することが多いため, 規則を追加するだけで対応できると考えられる.

4.3.1 誤抽出による弊害

SF の抽出という観点から日付・時間表現の誤抽出は他の SF に影響を与えることはないため, 大きな問題とはならない. 日本文の場合は誤抽出した語は固有名詞になり, 英文の場合は語抽出した語はノードへ編入する.

5. むすび

本論文では SF に基づく日英機械翻訳における日付・時間表現の抽出について提案した. 評価のために日英の日常会話文が中心のコーパスを使用した. 実験の結果, 日本文では適合率 96.7%, 再現率 98.2%, 英文では適合率 94.7%, 再現率 92.7%だった. この結果から提案手法により SF のための日付・時間表現を抽出できることが分かった. 抽出に失敗した表現には日付・時間表現が慣用句の中に現れる場合があり, 使用頻度の高い慣用句を除外する規則が不足していることが分かった. また今回, 日付・時間表現の抽出に主眼をおいているため, 前置詞までを考慮したことによる影響は検討できていない. 今後は使用頻度の高い慣用句を除外する規則の追加と前置詞を含めて変数化した影響を調査する予定である.

謝辞

本研究の一部は科学研究費補助金 (基盤研究 (B) 19300029, 17300065, 萌芽研究 17656128) の補助を受けて行った.

文献

[1] E. Sumita, Example-based machine translation using DP-matching between word sequences, ACL, 2001
[2] E. Sumita, H. Iida, and H. Kohiyama, Translating with Examples: A New Approach to Machine Translation, In Proceedings of 1st Theoretical and Methodological Issues in Machine Translation, 1990
[3] F. Ren, Super-function based machine translation, Language Engineering, Proceedings of JSCL and TsingHua University Press, pp.305–312, 1997
[4] M. Sasayama, F. Ren, S. Kuroiwa, Super-Function based Japanese English Machine Translation Experiment and Evaluation, Proceedings of the third International Conference on Information, pp.195–198, 2004
[5] R. D. Brown, Adding linguistic Knowledge to a Lexical Example-Based Translation System, Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation, pp.22–32, 1999
[6] R. Gangadharaiah, R. Brown, J. Carbonell, Spectral Clustering for Example Based Machine Translation, Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp.41–44, 2006
[7] K. McTAIT, Memory-based translation using translation Patterns, UMIST, 2001
[8] 北村美穂子, 松本裕治, 対訳コーパスを利用した翻訳規則の自動獲得, 情報処理学会論文誌, Vol.37 No.6, pp.1030–1040, 1996
[9] P.F. Brown, J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roosin, A statistical approach to machine translation, Computational linguistics Vol.16, No.2, 1990
[10] H. Watanabe, S. Kurohashi, and E. Aramaki, Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation, International Conference on Computational Linguistics(COLING), pp.906–912, 2000
[11] H. Kaji, Y. Kida, and Y. Morimoto, Learning translation templates from bilingual text, Proc. of COLING, pp.672–678, 1992
[12] ChaSen version 1.0 is officially released on 19 February 1997 by Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology
[13] E. Brill, Some advances in transformation-based part of speech tagging, Proceedings of AAAI, 1994
[14] EDR 電子化辞書仕様説明書, 通信総合研究所
[15] M. Sasayama, F. Ren, S. Kuroiwa, Automatic Extraction of Super-Function from corpus and experimental evaluation, Information-MFCSIT'06, pp.395–398, 2006
[16] 篠田義明, 科学技術論文に頻出する英語表現 2, 日興企画, 1993