語境界の差異を伴う形態素情報変換

下畑 光夫 隅田 英一郎
ATR 音声言語通信研究所
{mshimoha,sumita}@slt.atr.co.jp

1　はじめに

コーパス、辞書、形態素解析ツールなど様々な自然言語処理リソースが公開されているが、それらで採用されている品詞体系は異なっていることが多い。例えば、日本語のタグドコーパスとして、京大コーパス[1]、RWCコーパス[2]、EDRコーパスがよく知られているが、それぞれ異なった品詞体系を使用している。

品詞体系により品詞の集合、定義が異なるため、異なる体系の品詞情報をそのままの形で合わせて利用する事は困難である。異なる体系の品詞情報を一緒に用いるには体系変換が必要となる。

日本語は語境界が現れないので体系により語区切りが異なる場合も生じることがある。形態素情報[1]を異なる体系に変換する場合、　品詞とともに語境界の差異も変換する必要がある。我々は、語境界が一致する語の間での品詞情報変換の場合、前後2語の品詞を属性とすることで98.8%の精度で変換できることを既に報告した[3]。

本論文では、語境界の差異を伴う場合の形態素情報変換方法について述べる。語境界の差異を含む変換を扱いやすくするために、セグメントという単位を導入する。また、変換は、語彙化変換と一般変換の2つの方式を組み合わせて行う。語彙化変換は、出現頻度を元に個々のセグメントに特化した変換を行なう。一般変換は、セグメントの構成語の異なり頻度を元に、一部の構成語の表記条件を緩和した変換パタンで行なう。一般変換は新出語に対しても適用することができる。本手法をJUMAN体系(京大コーパス)とIPA体系(RWCコーパス)の間で適用した実験について報告する。

従来の日本語形態素情報の変換の研究[4][5]においても語境界が変動する場合の変換を考慮しているが、語個別に扱って新出語に適用できないことや、複数の語が変換により単一の語に統合される場合は扱えないなど、対応できる場合は制限されている。

[1]本論文では、変換する情報は品詞だけでなく語境界も含むため、以降では形態素情報と呼ぶ。

2　セグメント

語を単位として語境界の差異を伴う変換を行なうと、語長や構成語数が変化するために取り扱いが煩雑となる。そこで、変換の単位を語ではなくセグメントとする。セグメントは、元体系、目的体系で共通する語境界で分割された部分列と定義する。セグメントを単位とすると、語境界の変動はセグメント内で考えればよく、セグメントの語長は変換前後で変化しない。また、変換前後でセグメントの個数も変化しない。

図1に2体系の形態素情報が付与されたテキストを示す。IPA体系の方が語区切りが短単位となっており、「的に」、「して」、「きた」の語がIPA体系では短く分割されている。共通する語境界で分割することにより、5つのセグメントに分割されている。ここで、両体系で1つの語のみを含むセグメントをSS(Single-Single)セグメントと呼び、元体系で1語、目的言語で複数の語を含むセグメントをSM(Single-Multi)セグメントと呼ぶ。同様にして、MSセグメント、MMセグメントも定義される。図1ではセグメント1,3はSSセグメント 2,4,5はSMセグ

JUMAN 体系				
全国 名詞	的に 接尾辞	拡大 noun	して 動詞	きた 接尾辞
1	2	3	4	5
全国 名詞	的 名詞 | に 助詞	拡大 名詞	し 動詞 | て 助詞	き 動詞 | た 助動詞
IPA 体系				

図 1: 同一テキストに付与された 2 体系の形態素情報

メントである。体系間で語境界の変動を伴うのは、SM,MS,MM セグメントである。

3 語境界の変更を伴う変換方法

京大コーパス、RWC コーパスの共通テキスト部分を元に、両体系セグメントを構成語の個数を調べると表 1のようになる。語境界の差異を伴わない SS セグメントが最も比率が高く、全体の 85.8% を占めているが、それ以外の語境界に差異があるセグメントも合計で 14.2% を占めていることが分かる。

変換は、語彙化変換と一般変換の 2 つの方式を組み合わせて行なう。変換対象となるセグメントの頻度により変換方式が使い分けられる。以下の節で両変換の変換規則の生成方法ならびに生成された規則の実例について述べる。

変換は語彙化変換が優先して適用され、語彙化変換が適用されなかった語は一般変換が適用される。一般変換も適用されなかった語は、語境界の差異を伴う語と見なされないことになる。つまり、語境界の変化を伴うかどうかの判定も同時に行なっていることになる。

表 1: セグメント内の構成語数

分類	構成語数		個数	割合
	JUMAN	IPA		
SS	単数	単数	691,640	85.8
SM	単数	複数	92,367	11.5
MS	複数	単数	19,226	2.4
MM	複数	複数	2,451	0.3

3.1 語彙化変換

語彙化変換の変換規則の生成手順は以下の通りである。図 2に生成された語彙化規則を示す。

1. トレーニングデータから SM,MS,MM セグメントを集計し、頻度がしきい値を超えるセグメントを収集する。頻度のしきい値は 100 とした。
2. 収集された各セグメントについて、最も多く現れた目的体系のセグメントを抽出するし、それを変換結果とする規則を生成する。
3. 生成された規則をトレーニングデータに適用し、適合率が 50% を越える規則を変換用規則として抽出する。

適用条件 (JUMAN 体系)	変換結果 (IPA 体系)
して (動詞)	し (動詞)+ て (助詞)
である (判定詞)	で (助動詞) + ある (助動詞)
いた (接尾辞)	い (動詞) + た (助動詞)
れて (接尾辞)	れ (動詞) + て (助詞)

図 2: 生成された語彙化変換規則

3.2 一般変換

一般変換では、MS,SM セグメントについて複数語から構成される体系における語の一部を変項化する。MM セグメントは変項化する部分が両体系で交差することが考えられるため、対象としない。変換パタン生成手順は以下の通りである。

1. トレーニングコーパスから SM,MS セグメントを抽出する。複数語から構成される体系を元に、各構成語について表記情報を変項化したパタンを作成する。異なるセグメントから同一のパタンが生成されることがあるので、あるセグメントから生成された時にパタンの頻度を 1 加えることとし、各パタンごとに頻度を集計する。
2. (1) の中から頻度がしきい値を越える変項化パタンを抽出する。しきい値は、変項化した語について、機能語であれば重み 3、内容

図 3: 生成された一般変換パタン

適用条件	変換結果	適合率
＊である (形容詞)[2]	＊ (名詞) + で (助動詞) + ある (助動詞)	93
＊だった (形容詞)	＊ (名詞) + だっ (助動詞) + た (助動詞)	91
＊であり (形容詞)	＊ (名詞) + で (助動詞) + あり (助動詞)	91
＊続けて (動詞)	＊ (動詞) + 続け (動詞) + て (助詞)	100
＊的な (形容詞)	＊ (名詞) + 的 (名詞) + な (助詞)	88
＊よう (動詞)	＊ (動詞) + よう (助動詞)	98

語であれば重み 10 として、総和をとった値とした。

3. (2) で抽出されたパタンをトレーニングコーパスに適用し、適合率、再現率を算出する。50% を越えるパタンは変換パタンとして採用し、適合率を記録する。

生成された一般変換パタンを図 3に示す。複数の変換パタンが適用可能な場合は、適合率が最も高いパタンを適用する。

4 実験

JUMAN 体系 (京大コーパス) を元体系に、IPA 体系 (RWC コーパス) を目的体系にして変換実験を行なった。両コーパスでは毎日新聞 95 年の 2,928 記事 (39,065 文、923,305 セグメント) を共通して使用しており、変換前後の対応をとることができる。京大コーパスの形態素情報は人手によるチェックが入っているが、RWC コーパスは機械付与されただけである。このデータを二分し、一方をトレーニングデータ、もう一方を評価用データとした。変換する品詞は大分類とした。JUMAN 体系では 13 種類、IPA 体系では 10 種類の品詞が存在する。

4.1 語彙化変換

表 2に語彙化変換の変換精度を示す。語彙化変換はセグメント個別の規則であるため、後述の一般変換と比較すると精度が高い。

[2]"＊" は語の表記に対するワイルドカード、カッコ内は前接語の品詞を表す。

表 2: 語彙化変換の精度

規則数	55
適用事例	25,377
正解変換	24,210
変換精度	95.4%

4.2 一般変換

一般変換は一部の語を変項化しているために、新出語に対して正しく適用できると同時に、誤って適用される恐れが大きい。実験では、SM,MS セグメントの変換に分けて適合率、再現率の評価を行なった。また、語境界の変化を伴うかどうかの判定と変換パタンが適用されるかどうかと適用された上で変換結果が正しいかどうかも評価した。分類・変換精度を表 3に示す。SM セグメントの変換では、163 個の変換パタンが作成された。

表 3: SM セグメントの分類・変換精度

	正解語数	変換語数	個数
I	複数	複数	(正変換) 14,599 (総数) 15,518
II	単数	複数	1,231
III	複数	単数	10,661

表中、II,III は語境界の変動を伴うかを誤って

表 4: MS セグメントの分類・変換精度

	正解	変換	個数
I	単数	単数	(正変換) 93 (総数) 93
II	単数	複数	5,147
III	複数	単数	90
IV	複数	複数	1,204

判定した場合を表している。判定における適合率は $\frac{I}{I+II}$ = 92.7%、再現率は $\frac{I}{I+III}$ = 40.7% となる。変換精度は 53.3% となる。

また、MS セグメントでは 4 個の変換パタンが生成された。表 4に分類・変換精度を示す。判定の適合率は $\frac{I}{I+III}$ = 50.8%、再現率は $\frac{I}{I+II}$ = 1.8% である。IV は、元体系、目的体系の両方で複数語を含む場合であるが、この場合に対応するパタンは生成されないため、すべて変換エラーとなる。したがって、MS セグメントにおける変換精度は、1.4% となった。JUMAN 体系と THiMCO 体系を比較すると、全体的には JUMAN の方が語を長単位にする体系であるが、複合名詞では短単位となることがある。したがって、SM セグメント変換では対象となるセグメントも多く、機能語も含んだ一般性のあるパタンが多く抽出された。しかし、MS セグメント変換は事例が少なく、対照的な結果となった。

4.3 全体の変換精度

語境界の差異を伴う各変換について、適用事例数、正しく変換された事例数を表 5に示す。語境界の差異を伴う変換かどうかを判定する精度は 69.1%, 正しい変換結果を得られる精度は 65.6% であった。語境界の差異を伴う変換は、事例数が多く獲得できなかったことや、語の固有性に依存する部分が大きいことから、一般変換では十分な精度が得られなかった。語境界の差異がないセグメントと正しく判定された語は 342,980 個あった。この場合は 98.8% の精度で変換できる [3] ことを利用すると、すべての語に対する変換精度は 94.0% となる。

表 5: 混成変換における精度

	語彙化	一般	
		SM	MS
適用事例	25,377	27,410	6,534
正変換数	24,210	14,599	93
変換精度	95.4%	53.3%	1.4%

5 まとめ

本論文では、語境界の差異を伴う形態素情報の変換方法について述べた。語彙化変換と一般変換の 2 つの方式により、個別性の高い変換規則と一般性のある変換規則を生成し、両者を組み合わせることで 94.0% の変換精度を実現した。日本語のように語境界が明示されない言語で形態素情報を体系間で変換する場合、語境界の差異は避けられない問題であり、この問題の解決のための手法を示した。

参考文献

[1] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 第 3 回言語処理学会年次大会, pp. 115–118, 1997.

[2] 井佐原均. テキストコーパスの作成 – RWC, JEIDA, Orchid –. 第 11 回人工知能学会全国大会, pp. 54–57, 1997.

[3] 下畑光夫, 隅田英一郎. 形態素体系間の情報変換手法. 情処研報自然言語処理研究会, Vol. 2001, No. 9, pp. 157–162, 2001.

[4] 田代敏久, 森元逞. 形態素情報付きコーパスの再構成手法. 情報処理学会論文誌, Vol. 37, No. 1, pp. 13–22, 1996.

[5] K. Inui and H. Wakigawa. A pos-tag conversion algorithm for reusing corpora. In *Proceedings of NLPRS*, pp. 56–61, 1999.