自然言語処理 141－25
（2001．1．26）

形態素体系間の情報変換手法

下畑光夫　隅田英一郎
ATR 音声言語通信研究所

mshimoha@slt.atr.co.jp, sumita@slt.atr.co.jp

コーパスや辞書を初めとする自然言語リソースの多くには形態素情報が付与されている。しかし、これらの形態素情報には様々な体系が存在し、各体系で品詞の定義が異なっており、そのままでは一緒に使うことができない。
本論文では、異なる体系の形態素情報を変換する方法について述べる。変換方法は、語彙化変換と一般変換の2種類を組み合わせて行なう。語彙化変換は、頻度の高い語に対して適用され、語境界の変更を伴う変換が可能である。また、一般変換は頻度の低い語や新出語に対する変換に適用される。京大コーパス (JUMAN 体系) と RWC コーパス (THiMCO 体系) の間での変換の実験結果について報告する。

Morphological Information Conversion Between Different Systems

Mitsuo Shimohata, Eiichiro Sumita

ATR Spoken Language Translation Research Laboratories

Many linguistic resources, such as corpora and dictionaries, contain morphological information. There are many kinds of morphological information systems that differ in style and in definitions for part-of-speech. Therefore it is difficult to use them together as they are.
In this paper, we describe a method of converting morphological information between different systems. Conversion consists of two methods: lexicalized conversion and general conversion. Lexicalized conversion is applied to frequent words and enables conversion involving word boundary changes. General conversion can be applied to infrequent words and unseen words. We also describe the conversion experiment between the Kyodai corpus and the RWC corpus.

1 はじめに

品詞、活用形、辞書形などの形態素情報が付与されたタグドコーパスが多く公開されているが、それらのコーパスで採用されている品詞体系は様々である。京大コーパス [1], RWC コーパス [2], EDR コーパス [4], ATR 音声言語データベース [3] などがよく知られているが、それぞれ採用している品詞体系は異なっている。したがって、これらのコーパスの形態素情報をそのまま合わせて使用することができない。

本論文では、異なる体系間で形態素情報を変換する方法について述べる。変換は、語彙化変換と一般変換の２つの方式を用いている。語彙化変換は個別の変換対象語ごとの変換規則を利用する方法であり、語境界の変動を伴う変換も可能である。一般変換は品詞のみを参照する変換規則を利用する方法であり、語彙化変換と比較すると精度はやや低いが、新出語の変換を行なうことができる。両方式を頻度により使い分ける。

田代ら [6] は、基本的に変換対象語を品詞と表記で区別して、語個別に変換規則を生成している。表記が規則にとって未知の場合、語長を利用したり、同名称の品詞に対応すると見なすなどのヒューリスティックで対応している。また、乾ら [7] は、目的言語の連接関係、係り受け情報に基づいて変換規則適用のあいまい性解消を行なっている。

2 形態素情報変換における問題点

体系が異なると品詞の定義が異なってくるため、２つの体系で同じ名称の品詞があってもカバーする範囲が等しいとは限らないし、一方の体系にしか存在しない品詞もある。また、日本語などの語境界が明示されない言語では、どこで語を区切るかが体系によって異なる。体系間の変換では、この２つの差異を適切に変換することが求められる。以下で、両者の差異の変換について詳しく述べる。

2.1 品詞のあいまい性

表 1に、京大コーパスと RWC コーパスにおける品詞の対応を示す。両体系とも、名詞、動詞、形容詞、動詞、助動詞、助詞、連体詞、接続詞、接頭辞、感動詞が定義されている。指示詞、判定詞、接尾辞は JUMAN 体系にのみ定義されている品詞である。THiMCO からみた連体詞を除いて、同名称の品詞に対応する場合が最も多いが、別名称の品詞に対応している事例も少なからず存在する。THiMCO 体系の連体詞は JUMAN 体系では連体詞ではなく、指示詞に最も多く対応している。この他、両体系において助詞が比較的近い概念を表していることや、JUMAN 体系の接尾辞が THiMCO 体系においては名詞性接尾辞、動詞、助動詞などの広い概念を含んでいることが分かる。

本論文では、このように多様性を持つ品詞間の対応に対して、変換対象となる語の形態素情報と近接する語の形態素情報を属性として利用し、変換を行なう。事前知識やヒューリスティックは利用しない。

2.2 語境界の差異

日本語は語境界が表記に明示されておらず、語区切りの基準が体系により様々である。変換元、変換先における語長の差異を語単位で取り扱うのは困難である。それは、変換先の語長が長くなる場合に新たな文字を補わねばならないことや、テキスト全体において語長変化の整合性がとれていなければならないためである。

そこで、本方法では変換の単位を “語” でなく、“セグメント” とした。セグメントとは、変換元、変換先体系の両方で共通する語境界で分割した文字列のことである。セグメントに分割された文の例を図 1に示す。THiMCO 体系では、2,4,5 番目のセグメントは２語を含んでいる。セグメントを単位とすることで、語境界の差異はセグメント内だけで考えればよいことになる。セグメント内で語境界は変更することはあるが、セグメントそのものは変換前後で長さは変わらない。

JUMAN 体系

全国	的に	拡大	して	きた
名詞	接尾辞	名詞	動詞	接尾辞
1	2	3	4	5

THiMCO 体系

全国	的 - に	拡大	し - て	き - た
名詞	名詞 - 助詞	名詞	動詞 - 助詞	動詞 - 助動詞

図 1: セグメント分割

表 1: 京大コーパスと RWC コーパスの間での品詞対応

京大 (JUMAN)	RWC (THiMCO)									
	名詞	動詞	形容詞	助動詞	助詞	助動詞	連体詞	接続詞	接頭辞	感動詞
名詞	305,088	637	23	283	215	13	18	10	1,942	-
動詞	508	47,978	57	146	81	28	45	34	5	1
形容詞	886	9	7,881	298	4	33	315	1	1	-
助動詞	3,006	19	9	7,558	10	-	4	386	37	-
助詞	203	69	-	56	232,881	499	-	200	-	-
助動詞	774	2	3	1	243	1,614	-	-	-	-
連体詞	1	24	3	2	1	1	713	-	3	-
接続詞	5	4	-	22	-	-	-	3,265	-	1
接頭辞	351	2	-	-	-	-	1,027	-	4,023	-
感動詞	32	4	-	2	-	-	-	29	-	10
指示詞	2,584	3	-	802	-	-	4,580	23	-	-
判定詞	4	6	-	-	997	3,958	-	-	-	-
接尾辞	30,935	14,563	1392	-	144	6,935	-	-	187	-

3 変換方式

表 1に示したように、品詞ごとの対応を概観すると、多様な関係が存在している。しかし、個々のセグメントごとに対応を見ると、多くのセグメントは取りうる変換先は 2 個以下であり、個別に変換を行なう方が精度が高い。そこで、ある程度の事例数があるセグメントについては個々に変換規則を生成し、事例が少ないセグメントは全体で一つの変換規則を生成する。両者の切り分けは頻度のしきい値によって決定される。しきい値より頻度が高いセグメントは個別の変換規則による変換 (語彙化変換) を、また頻度が低いセグメントは表記を問わない変換 (一般変換) を適用する。

変換規則の生成手順を図 2に示す。学習データとして、図 1のような 2 つの体系の形態素情報を持つテキストを使用する。テキストを両体系で共通する語境界で分割し、セグメントを切り出し、集計する。頻度がしきい値以上のセグメントは語彙化変換の対象となり、それ以外のセグメントは一般変換の対象となる。セグメントは、表記と品詞を合わせて区別される。例えば、形容詞の “ない” と動詞の “ない” は区別される。

各セグメントについて、変換元体系の情報からセグメント自身の品詞と活用ならびに前後 2 語の品詞を収集し、属性とする。また、変換先体系の情報から変換先品詞を取り出し、分類すべきクラスとする。これらの属性とクラスにより学習事例が構成される。

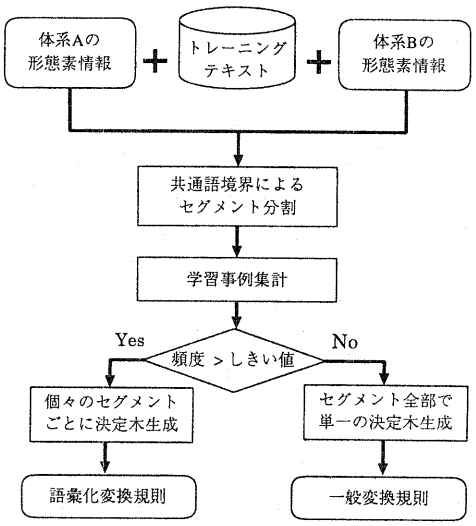

図 2: 変換規則の生成手順

学習事例から変換規則は決定木学習で獲得される。決定木学習ツールとして C5.0 [5] を使用している。変換規則を用いて変換を行なう場合、語彙化変換が優先して適用される。語彙化変換で変換できなかったセグメントが一般変換により変換される。

3.1 語彙化変換

語彙化変換では、個々のセグメントごとに変換規則が生成される。図 1のセグメント 3 から学習事例

属性				クラス
-2 語	-1 語	+1 語	+2 語	変換先
接尾辞	名詞	接尾辞	文末	し (動詞) + て (助詞)

図 3: “して (動詞)” の学習事例

属性					クラス
-2 語	-1 語	対象語	+1 語	+2 語	変換先
名詞	接尾辞	名詞	動詞	名詞	名詞

図 4: “拡大 (名詞)” からの学習事例

を生成すると図 3のようになる。変換結果は表記情報を保持しているため、セグメント内で語境界が変化する情報も変換結果に入れることができる。

3.2 一般変換

頻度の低いセグメントはまとめられ、属性として品詞を用いて単一の決定木が生成される。図 1 のセグメント 3 “拡大” から生成した学習事例を図 4に示す。生成された変換規則には表記的な条件がないため、新出語 (未知の表記であるが、品詞は判明している語) の変換も行なうことができる。この方式は変換先、変換元において一つの語しか含まないセグメントにのみ適用することができる。

4 実験

京大コーパス (JUMAN 体系) と RWC コーパス (THiMCO 体系) を用いて、JUMAN 体系から THiMCO 体系への変換実験を行なった。2-fold のクロスバリデーションで評価した。両コーパスとも、毎日新聞 1995 年の記事テキストに形態素情報を付与しており、共通する 2,929 記事を利用した。京大コーパスの形態素情報は人手による修正が行なわれているが、RWC コーパスは機械付与の状態で人手修正は入っていない。実験に用いたデータの仕様を表 2に示す。表中、単一形態素セグメントとは、構成語が 1 つしかないセグメントのことを指し、変換前後で語長が変化しない語のことを表している。全セグメントの 87.5% を占めている。

語彙化変換と一般変換は、セグメントの頻度によって使い分けられる。しきい値を下げると語彙化変換がカバーするセグメントは増加するが、少ない事例

表 2: 実験に用いたデータ

記事数	2,928
文数	39,065
セグメント数	923,305
単一形態素セグメント	807,961

	京大	RWC
のべ形態素数	946,529	1,026,892
異なり形態素数	42,698	34,968

表 3: 頻度しきい値による語彙化変換の精度

頻度	種類数	事例数	誤り数	精度
～ 1,000	10	105,522	527	99.5%
999 ～ 500	11	7,934	190	97.6%
499 ～ 100	80	16,796	567	96.6%
99 ～ 50	92	6,161	386	93.7%
49 ～ 10	358	7,664	888	88.4%

から変換規則を生成するセグメントでは精度が低下する恐れがある。複数の変換先品詞を持つセグメントを対象に、語彙化変換で変換するための頻度のしきい値を色々な場合に区切った場合の変換精度を表 3に示す。頻度 100 以下では、後述の一般変換より精度が低下しており、高い精度を実現するという効果が失われている。そこで、本実験ではしきい値を 100 に設定する。

セグメントを頻度により二分し、語彙化変換、一般変換を行ない、品詞別に集計した結果を表 4に示す。トータルでは語彙化変換の方が若干よい精度を示している。一般変換では品詞ごとの精度の差が大きいが、語彙化変換では安定している。接頭辞、接続詞、助詞などが一般変換において精度が低いが、これらの品詞は、個別の独自性が強い語が多く含まれており、一般変換の単一の決定木ではそれらの語の独自性を十分に表現しきれなかったということを示しているといえる。語彙化変換では、語境界の変

表 4: 語彙化変換・一般変換の品詞別精度

	語彙化変換		一般変換	
	変換数	適合率	変換数	適合率
名詞	117,120	99.8	229,373	99.1
形容詞	4,014	99.4	5,283	92.6
接頭辞	3,524	94.7	1,272	43.4
感動詞	0	-	4	8.3
連体詞	5,620	99.6	1,131	90.5
接続詞	2,234	99.1	1,185	64.2
助詞	238,038	99.9	1,610	71.6
副詞	2,038	97.8	7,901	86.3
動詞	34,533	99.7	28,790	97.1
助動詞	11,626	97.2	1,106	94.1
合計	418,747	99.8	277,655	97.4

動を伴うセグメントも対象とすることができ、実験では 34 種類、事例数 8,790 に適用された。

また、一般変換で生成された決定木の一部を図 5に示す。これより、名詞、助詞、動詞などは、両体系でカバーする範囲に多少の差異はあっても全体としてみた場合にはほぼ同じ概念を表しているといえる。JUMAN 体系では接尾辞は広い概念を含んでいるが、これも直前語の品詞をみることで判別できることを表している。

語彙化変換のカバー率を図 6に示す。ハッチ部分は、語彙化変換により変換された割合を表している。語彙化変換は、助詞、助動詞に高い比率で作用している。これは両品詞の語は種類が少なくて頻度が高い性質を持っているためである。一方、名詞、副詞の各語は低い頻度の語が多いため、比率としては大きくならない。また感動詞は新聞ではほとんどで出現しないため、語彙化変換の対象とならなかった。全体では 418,747 個のセグメントが語彙化変換で変換され、277,655 個のセグメントが一般変換で変換された。語彙化変換により 60.1% のセグメントが変換されたことになる。両変換を混成した変換品詞と実際の品詞の対応を表 5に示す。変換精度は 98.8% である。

変換元の条件			変換先品詞
対象品詞	=	名詞	名詞
対象品詞	=	接続詞	接続詞
対象品詞	=	連体詞	連体詞
対象品詞	=	助詞	助詞
対象品詞	=	動詞	動詞
対象品詞	=	接尾辞	
直前語品詞	=	名詞	名詞
直前語品詞	=	副詞	名詞
直前語品詞	=	形容詞	動詞
直前語品詞	=	助動詞	動詞

図 5: JUMAN 体系から THiMCO 体系への一般変換規則

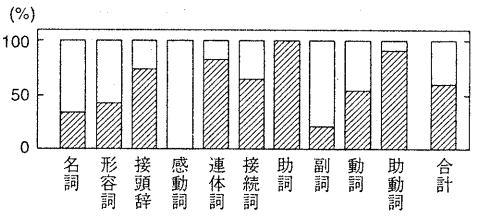

図 6: 変換全体に対する語彙化変換のカバー率

5 考察

本実験では、語彙化変換と一般変換を組み合わせて 98.8% の精度を実現することができた。この精度は JUMAN 体系から THiMCO 体系への変換においてのものであり、逆方向の変換や他の品詞体系への変換では精度が変わる可能性がある。JUMAN 体系は THiMCO 体系と比較すると品詞の種類が多いので、その点では有利な変換であるといえる。変換元、変換先の体系を色々変えた実験については今後行ないたい。

本方法では、複数の語で構成され、かつ頻度が低いセグメントに対する変換に対応できていない。このようなセグメントは、全セグメントの 11.5% を占めている。語彙化変換を行なうためのしきい値を下げれば、これらのセグメントを語彙化変換に多く取り込むことができ、変換できないセグメントを減少させることができる。しかし、表 3に示したようにしきい値を下げると、精度の悪い変換が混入して

表 5: 混成変換における変換精度

目的体系の付与品詞	変換処理による付与品詞									
	名詞	形容詞	接頭辞	感動詞	連体詞	接続詞	助詞	助動詞	動詞	助動詞
名詞	342,808	166	268	1	16	27	210	1,236	411	3
形容詞	103	8,945	-	-	3	-	-	8	281	29
接頭辞	1,596	-	4,525	-	45	-	-	29	2	-
感動詞	1	-	-	1	-	8	-	-	2	-
連体詞	40	24	-	-	6,575	-	-	18	43	-
接続詞	20	-	-	2	5	3,318	174	421	7	-
助詞	396	3	-	-	1	28	238,882	17	71	4
副詞	745	114	1	-	79	29	28	8,188	34	-
動詞	763	11	2	-	26	9	68	20	62,413	15
助動詞	21	34	-	-	1	-	286	2	59	12,681

しまうという問題がある。また、新出のセグメントには対応できない。他の方法として、セグメントを構成する語のいくつかの表記条件を緩和することが考えられる。この方法は、語境界の変動を伴う新出セグメントにも適応できるという根本的な解決策でもあり、今後検討したい。

6 まとめ

本論文では、異なる体系間で形態素情報を変換する方法について述べた。語彙化変換と一般変換の2種類の方式を組み合わせ、98.8% という高い精度を達成することができた。語彙化変換により精度の高い変換と語境界の変更を伴う変換を実現し、一般変換により頻度の低い語ならびに新出語の変換を実現した。

参考文献

[1] 黒橋、長尾:“京大テキストコーパス・プロジェクト”, 人工知能学会, 第 11 回全国大会,(1997).

[2] 井佐原:“テキストコーパスの作成”, 人工知能学会, 第 11 回全国大会,(1997).

[3] 竹沢 :“音声コーパスの構築と利用”, 人工知能学会, 第 11 回全国大会,(1997).

[4] EDR (Japan Electronic Dictionary Research Institute Ltd.): The EDR Electronic Dictionary Technical Guide. In *EDR Technical Report* (in Japanese). (1995)

[5] http://www.rulequest.com/see5-info.html

[6] 田代、森元:“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,(1996).

[7] Inui K, Wakigawa H: A POS-tag Conversion Algorithm for Reusing Corpora. In *Proceedings of NLPRS.* (1999)