The ATR/Lancaster General – American – English Treebank

柏岡 秀紀 , Stephen G. Eubank , Ezra W. Black
ATR音声翻訳通信研究所

1 はじめに

計算機による自然言語処理には，内省により得られる知識や大規模な言語データの分析から得られる知識が用いられている．

内省による知識は，個々人による揺れがあり，また定量的な分析が困難である．そこで，処理対象となるドメインのテキストで知識を洗練することにより，揺れている知識を検証し，精度の向上を図る．

これに対して，大規模な言語データからは，客観的で定量的な知識を得ることができる．特に，予め分析された大量のデータからは，より有効な知識を得ることができる．

どちらの場合も言語データの利用が重要になる．近年，言語データから得られる統計的知識や用例を用いた処理技術を活用した研究が盛んである [1, 2, 3, 4, 5, 6, 7, 8].

日本語の言語データに関しては，徐々に整備され始めている．分析ずみのデータとして，ATRの音声言語データベースやEDRのコーパスがある．最近では，大量の新聞記事データを比較的容易に入手できるようになってきている．また，一般の辞書や小説がCD-ROMの形態で販売されており，利用できるデータが増えつつある．

英語に関しては，LDCなどにより整備されており，Brounコーパスやペンシルベニア大学のツリーバンク，TIPSTARのデータなど大規模な言語データを手に入れることができる．

現在，英語の大規模な分析ずみ言語データベースが普及している．それに付与されている構文構造の情報は，多くの場合，“skelton parsing”による構文構造の情報が付与されている [9, 10, 11]. このような構文構造の骨組みの情報から得られる知識を構文解析に用いる場合，2節で述べるような問題が生じる [12]. それらの問題は，十分な構文構造の情報が付与された言語データベースから知識を得ることで解消できると思われる．我々は，詳細な情報を付与した大規模な言語データベースの開発を行ない，ATR/Lancaster General English Treebank（以後，ATR-E-TBとする）を作成した．本稿では，このATR-E-TBの内容について報告する．

2 特徴

大規模な分析ずみ言語データベースを構築するにあたり，対象とする文書の選択，および分析に使用する文法体系は，重要なポイントである．

対象とする文書

ATR-E-TBでは，特定のドメインに限定することなく，ドメインに依存しない言語を取り扱うことにした．そのTBにより，統一的な文法体系のもとでドメインに偏らない言語知識を得ることを目指した．各ドメインに対しては，この知識を改変することにより対応できると考えている．

対象とした文書は，

1) インターネットから得た文書
2) OCRにより得た文書
3) ベンダーから購入した文書

から構成される．主として1990年代に標準的なアメリカ英語で書かれた文書を収集した．その際に，複数の分類(文書の長さ，主題，スタイル，作者の立場等)を行なった．以下に，ATR-E-TBに含まれる文書のタイトル例を示す．

- Empire Szechuan Flier (Chinese take–out food)
- Catalog of Guitar Dealer
- UN Charter: Chapters 1–5
- Airplane Exit–Row Seating: Passenger Information Sheet
- Bicycles: How To Trackstand
- Government: US Goals at G7
- Shoe Store Sale Flier
- Hair–Loss Remedy Brochure
- Cancer: Ewing's Sarcoma Patient Information

収集した文章には，Captain John SmithによるPlymouth Plantation(1600年代)やBenjamin Franklin(1700年代)により書かれた文書や，イギリス英語，オーストラリア英語，カナダ英語などアメリカ英語以外で書かれた文書もある程度含まれている．

文法体系

従来，大規模な言語データベースの作成手法として，“skelton parsing”による手法が提案されている．しかし，“skelton parsing”による言語データベースは，部分的で，相対的に概略だけの構文構造の情報を提供するに留まる．この種の言語データベースを統計的構文解析のトレーニングに用いた場合には，以下のような問題が生じる．例えば，トレーニング用のデータベース作成に用いた文法よりトレーニングしようとする解析用の文法規則が詳細な場合，解析用の文法は，十分なトレーニングがなされず，性能を出すことができない．また，解析用の文法が過適合してしまう場合もあり，誤ったトレーニング結果を出してしまう．トレーニング用のデータベース作成に用いた文法に確率を付与する場合でも，情報が欠落しているため，同様な問題が生じる．

我々は，構文構造に関する十分な文法情報を付与された言語データベースを用いることにより解消できると考えた．ATR-E-TBでは詳細な文法を用いて，十分な文法情報を付与することを目指した．他の解決法として，文法を用いない解析手法も提案されている [6, 7, 8]. 当然，ATR-E-TBは，文法を用いない手法にも使うことができる．

また，品詞セットに意味カテゴリを含めた．これにより得られる意味的知識による実用的な処理への効果も検証したい．

3 構成

ATR-E-TBは，総語数が約70万語，総語彙数が約5万語であり，約1,000の文書からなる．文書は，30語程度から3600語程度の長さであり，1000語を越える文書は，全体の20%前後である．

文書の分類

各文書には，複数の特徴による分類がなされている．表1に，その例を示す．

表1: 文書の分類に用いた特徴の例

視点	例
論調 (TONE)	friendly
文体 (STYLE)	dense
レベル (LINGUISTIC.LEVEL)	literary
分野 (POINT.OF.VIEW)	technical
種別 (PHYSICAL.DESCRIPTION)	guide
著者の出身 (GEOGRAPHICAL. BACKGROUND. OF. AUTHER)	American South
...	

以上のような特徴に関する情報は，各文書を読んだ読者が判断し与えている．また，各々の特徴の値を一つに絞ることができない場合や，不明なものもある．以下に示す割合は，その目安としてとらえた値である．

収集されている文書の90%程度は，アメリカ出身者により書かれたものである．また，第三者的立場で書かれている文書が80%程度，残り20%のほとんどが，一人称の立場で，書かれたものであった．さらに，標準的，形式的，口語的，文学的な文書の割合は，それぞれ70%，15%，8%，4%程度である．

これら以外に，より一般的な視点から，タイトルや著者の視点，文書の入手先などの値も付与されている．

構成ファイル

各文書は，SGMLを用いて表現される以下の4つのファイルにより構成した．

1) Original File
   オリジナルのテキストのファイル．
2) Hor File
   オリジナルのテキストに対して，単語数，行数などの情報と，段落情報を付与されたファイル．
3) Output File
   単語分割，文分割を行ない，各文の分析結果とともに文の数や使用した文法規則数の情報を付与したファイル．
4) Queries File
   分析時に，修正されたspellの情報や，分析に失敗した原因となる部分とその理由等の情報からなるファイル

分析された文の例を，図1に示す．語(w)とその品詞(t)は，“w_t”として表現している．適用された文法規則が明確にわかるように，その規則名をラベルとしたブラケットにより，規則の適用範囲を囲んでいる．このラベルつきのブラケットにより構文構造(木構造)が表現される．

4 文法

ATR-E-TBでは，前節で示した“skelton parsing”による言語データベースでの問題を解消するために，詳細な文法体系を用い，構文構造に関する十分な文法情報を付与している．ATR-E-TBには，ATRで開発した素性構造つき文脈自由文法を用いている．

文法の変更は，言語データベースの統一を崩し，利用価値を著しく低下させる．そのため，言語データベースの構築には，成熟した文法が必要とされる．大量のテキストを解析し洗練されているIBMや他のシステムで用いられている文法を参考にした．この文法は，主にIBM English Grammar[5]に見習って開発した．

品詞セット

品詞セットは，意味カテゴリを含め，おおよそ2200品詞に分類されている．意味カテゴリを除いた場合，165品詞に分類される．この品詞セットは，IBMの品詞セットより，UCRELにより開発されたClawsシステムの品詞セット(179品詞)に類似している．例えば，“mnemonic”の概念は，ATRの文法には応用されていない．

Clawsのシステムとの類似点として，ATRの品詞体系でも，“ditto tags”と呼ばれる品詞体系を用いていることがあげられる．これは，複数の語による表現に対する品詞付与を目的としている．例えば，“will o' the wisp”は，4語からなる単数の普通名詞として扱わる．各単語には，品詞セットの品詞(この例では，単数の普通名詞“NN1”)の後に構成する単語数(この例では，4)とその単語の位置(この例の“the”であれば3)を付与し，品詞(この例の“the”では，“NN1 43”)が与えられる．

また，Clawsとの大きな違いとして，“covering”を例としてその差異をしめす．“wall covering”として使われている場合，これは，“-ing”で終る名詞である，一方，“the covering of all bets”として使われている場合，これは，動名詞である．Clawsのシステムでは，これらは共に“NN1”(単数の普通名詞)として扱われるが，ATRの品詞セットでは，後者の場合，品詞に動名詞として“NVVG”を取り入れた．

品詞の意味カテゴリ

意味カテゴリは，名詞，形容詞，形容動詞に付随する42カテゴリと，動詞，準動詞に付随する29カテゴリがある．両者に，重複しているカテゴリもある．これらの意味カテゴリは，ドメインに関わらず，標準的なアメリカ英語の文書を対象として，ATRで開発したものである．このカテゴリ分類は，ATRで2人が6カ月間，その後Lancasterで5人が4カ月間品詞付与を行ないながら洗練し，その正当性を検証した．この意味カテゴリ分類は，より細かくすることもより粗くすることもできる．意味解析や実際の応用に当たっては，部分的により細かな分類が必要となったり，

解析文: It has meant great savings , both in time & gas ! ”

```
<S id="20" count=13>
<HIGH rendition ="italic">
[start [sprpd1 [sprime4 [sd1 [nbar6 It_PPH1 nbar6]
[vbar2 [o8 has_VHZ o8] [v2 meant_VVNMEAN [nbar12 [j1 great_JJDEGREE j1]
[n1a savings_NN2MONEY n1a] nbar12] v2] vbar2] sd1]
[iebar2 ,_, [i1e [pr1 [rmod1 [r2 both_RRCONCESSIVE r2] rmod1]
[p1 in_IIIN [coord1 [nbar1 [n1a time_NN1TIME n1a] nbar1]
[coord3 [cc3 [cc1 &_CCAMP cc1] cc3]
[nbar1 [n1a gas_NN1SUBSTANCE n1a] nbar1] coord3] coord1] p1] pr1] i1e]
iebar2] sprime4] [rand3 !_! "_"R rand3] sprpd1] start]
</HIGH>
</S>
```

図 1: *ATR-E-TB* の解析文の例

もっと粗い分類で良い部分があるとおもわれる. しかし, ここで与えた分類は, 多くの文書の分析を通じて, 洗練したものであり, 多くの人が納得できるカテゴリ分類であると考えている.

規則と素性

ATR-E-TB の作成に用いた文法は, 規則数が約 1100, 素性が 67 ある. これに対して, IBM の文法は, 規則数が 750, 素性が 40 である. つまり, ATR-E-TB で用いている素性は, IBM の文法の 1.5 倍強あり, 非常に些細と思われる情報も含まれている. その有効性は, 処理の目的や対象とするドメインにより変わり, その些細な情報が構文の曖昧性の解消に有効な場合もある. ATR-E-TB で用いている素性の例を表 2に示す.

表 2: 素性の例

素性	値
pos	v, n, s, determin, punct, ccxx, csxx ...
case	subject_case, object_case, reflexive ...
noun_group	has_j, numeric, letter, has_r, ing
wh_type	n, r, j, p

5 他の言語データとの比較

ATR-E-TB の語彙に関して, 特定のドメインを対象として集められた言語データとの比較を行なった[1]. その結果を示す.

以下の二つの言語データベースを比較の対象とした.

1) UPNN-WSJ
   ペンシルベニア大学のツリーバンク[2]の中の wall street jounal を対象としたデータ
2) AP88
   1988 年の AP の記事[3]を集めたデータ

[1]比較には, ATR-E-TB の約 560,000 語を用いた.
[2]LDC から購入
[3]LDC から購入した TIPSTAR に含まれているデータを使用

総語彙数の比較を表 3に示す[4]. これらのデータ間の共通の語彙数は, 表 4のようになっている. さらに, 各データにおいて, 総語数に対して共通の語彙が占める割合を表 5に示す.

表 3: 言語データベースの語彙の比較

言語データ	語彙数	総語数
ATR-E-TB	36	560
UPENN-WSJ	45	1,345
AP88	244	40,155

(単位: 千語)

表 4: 二つのデータの共通の語彙数

データセット	共通の 語彙数
ATR-E-TB & UPENN-WSJ	22
ATR-E-TB & AP88	29.5
UPENN-WSJ & AP88	38

(単位: 千語)

表 5: 共通の語彙が占める割合

共通語彙 \ データ	ATR-E-TB	UPENN-WSJ	AP88
ATR-E-TB & UPENN-WSJ	92	94	—
ATR-E-TB & AP88	97	—	92
AP88 & UPENN-WSJ	—	98	95

(単位:%)

平均的な文の長さを 20 語とした場合, 未知語の割合が, 1/20(5%) を越えていると, 未知語を含む文の割合が高くなる. 未知語に対応していないシステムでは, ほとんどの文が解析できないことになる.

[4]語彙を比較する時に, 表記の異る語は別の語彙とし, 数値は, 全てまとめて一つの語彙として扱った.

AP88の文を解析するために，ATR-E-TBをトレーニングに使った場合，トレーニングから得られる情報以外に辞書を用いなければ，大半の文の解析に未知語処理が必要となる．UPENN-WSJをトレーニングに使った場合も，ほぼ，同じと考えられる．

ATR-E-TBの語彙がカバーする割合は，ATR-E-TBの総語数がUPENN-WSJの半分弱であり，対象とする文書が特定のドメインに限っていないことなどを考慮する必要もあるが，高い値とはいえない．また，データの量を増加することで，未知語の割合が1%となっても，文単位の処理では，20%の文が未知語を含むことになる．これらの言語データベースから得られる情報以外に，何らかの未知語処理の技術が必要と考えられる．

6 おわりに

本稿では，ATR-E-TBの特徴，構成，および分析するために用いた文法について述べた．ATR-E-TBで用いた文法は，詳細な素性，品詞を持っている．

当然，文法体系をどれくらい詳細に記述すべきかという問題がある．これは，処理したい問題や対象，あるいは，処理手法により，一概にどれがよいということはできない．詳細な体系であれば，解析処理のための豊富で効果的な知識を得ることができ，実際の解析処理への活用が期待できる．逆に，詳細であるほど，個々の情報が現れる頻度が小さくなる．そのため，信頼できる知識を得るには，より大規模の言語データベースが必要となる．

そこで，ATR-E-TBの構築には，現状で必要と思われるもっとも詳細な情報を付与することを目的として，本論で述べた文法を用いた．したがって，ATR-E-TBは，粗い体系にも活用することができる．それにより，具体例に対する文法規則の確率の付与，意味情報の統計等を利用した多くの実用的な処理へ有効に活用できると考えられる．

今後，ATRで開発している統計的構文解析システムに，ATR-E-TBから得られる知識の活用を試み，その知識の有効性を判断し，ATR-E-TBを検証する．それとともに，ドメインを旅行に関する対話に限定し，ATR-E-TBの規模拡張を行ない，翻訳処理への活用を模索する．

また，ATR-E-TBとUPENN-WSJ，AP88の言語データに現れる語彙の比較を行なった．これにより，ATR-E-TBから得られる情報を構文解析等に利用する場合には，未知語に対する処理が重要と思われる．

参考文献

[1] E. Black, F. Jelinek, J. Lafferty, R. Mercer, S. Roukos: "Decision tree models applied to the labelling of text with parts-of-speech", *Proc. of DARPA Speech and Natural Language Workshop*, (1992).

[2] E. Brill: "Some Advances in Transformation-Based Part of Speech Tagging", *Proc. of the Twelfth National Conference on Artificial Intelligence*, pp.722-727, (1994).

[3] B. Merialdo: "Tagging English Text with a Probabilistic Model", *Computational Linguistics*, Vol.20, No,2, pp.155-171, (1994).

[4] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci: "Coping with Ambiguity and Unknown Words through Probabilistic Models", *Computational Linguistics*, Vol.19, No.2, pp.359-382, (1993).

[5] E. Black, R. Garside, and G. Leech, Editors: "*Statistically-Driven Computer Grammars Of English: The IBM/Lancaster Approach*", Rodopi Editions, Amsterdam,(1993).

[6] E. Brill: "Automatic grammar induction and parsing free text: A Transformation-based approach", *Proc. of 31st Annual Meeting of the Association for Computational Linguistics*, (1993).

[7] F. Jelinek, J. Lafferty, D. Magerman, R. Mercer, A. Ratnaparkhi, S. Roukos: "Decision Tree Parsing using a Hidden Derivation Model", *Proc. of ARPA Workshop on Human Language Technology*, pp.260-265, (1994).

[8] D. M. Magerman: "Statistical Decision-Tree Models for Parsing", *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics*, pp.276-283, (1995).

[9] E. Eyes and G. Leech: "Syntactic Annotation: Linguistic Aspects of Grammatical Tagging and Skeleton Parsing", Chapter 3 of Black et. al., (1993).

[10] R. Garside and A. McEnery: "Treebanking: The Compilation of a Corpus of Skeleton-Parsed Sentences", Chapter 2 of Black et. al., (1993).

[11] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz: "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, Vol.19, No.2, pp.313-330, (1993).

[12] E. Black: "An experiment in customizing the Lancaster Treebank" In Oostdijk and de Haan, pp.159-168, (1994).