自然言語処理 127－22
（1998. 9. 18）

日本語形態素・構文解析システムJEMONIの開発と評価について

石間 衛　藤井 敦　石川徹也

図書館情報大学
〒305-8550 茨城県つくば市春日 1-2
{ishima, fujii, ishikawa}@ulis.ac.jp

本論文では、我々が開発した日本語形態素・構文解析システム JEMONI を紹介する。JEMONI は主に情報検索システムで、検索質問や検索対象文書を解析することを目的に開発している。JEMONI と既存の形態素・構文解析システムとを 1) 名詞連続部分の推定精度、2) 係り受け先の解析精度の 2 点で比較評価した。また、評価の 1 手法である voting 法を提案する。voting 法とは、複数のシステムの共通解を評価における正解データとする方式であり、評価用コーパスを自動的に生成できるという特長を持つ。

Japanese Morphological and Syntactic Analyzer 'JEMONI' and its Evaluation

Mamoru Ishima　Atsushi Fujii　Tetsuya Ishikawa

University of Library and Information Science
1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
{ishima, fujii, ishikawa}@ulis.ac.jp

This paper proposes a Japanese morphological and syntactic analyzer 'JEMONI', which we currently use to analyze queries and documents for information retrieval systems. We evaluated JEMONI with other existing tools in terms of (a) the accuracy of extracting noun sequences from corpora, and (b) the accuracy of syntactic analysis. Since manual annotation for large-sized test corpora is still expensive, we propose an automatic evaluation method, in which we collect annotated corpora based on the degree of agreement among different analyzers.

1 はじめに

形態素・構文解析は、自然言語処理の基礎となる技術である。これまでに多くの研究が行われ、現在いくつかの実用システム ([1, 2, 3] 等) が公開されている。本論文では、我々が開発した形態素・構文解析システム JEMONI を紹介し、公開システムとの比較評価を行なう。

JEMONI は主に情報検索システムで、検索質問や検索文書を解析することを目的に開発している。我々は実用的な検索システムでは、高速、高精度に検索質問や検索文書を解析することが必要と考える。そこで JEMONI は形態素辞書に固有・普通名詞を持たない。固有・普通名詞を持たなければ、辞書容量の低下による解析速度の向上につながる。しかし他方において、複合名詞 (名詞連続) が分割されない問題がある。

JEMONI と既存の形態素・構文解析システムとを名詞連続部分の推定精度、係り受け先の解析精度の 2 点から比較評価した。

さらに本論文では、評価の一手法である投票方式 (voting 法) を提案する。voting 法とは、複数のシステムの共通解を正解データする方式であり、評価用コーパスを自動的に作成できるという特長を持つ。

以下、2 で JEMONI を紹介し、3 で他の形態素・構文解析システムとの比較評価を示す。4 で voting 法について説明する。

2 形態素・構文解析システム JEMONI

2.1 概要

JEMONI のシステム構成は図 1 の通りである。まず形態素辞書と連接規則を利用して、入力の「形態素解析」を行なう。形態素辞書は形態素を記述し、連接規則は連接可能な形態素の組を記述する。「形態素の接続」は形態素解析結果中の複数の形態素を 1 つにまとめる。「構文解析」は形態素の接続結果に対して、係り受け規則を利用して行なう。構文解析結果はユーザが指示した解析単位で出力される。

以下、2.2 で辞書、規則の作成方法を示し、2.3 でその構成を示す。2.4 で処理アルゴリズムを示す。

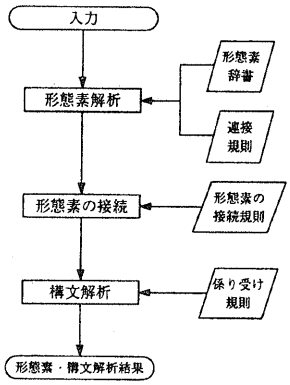

図 1: JEMONI のシステム構成

2.2 辞書、規則の作成

2.2.1 形態素辞書

形態素辞書を作成するために、以下の手順で既存の品詞情報付きの形態素データを抽出した。

i) 形態素データの抽出

CD-毎日新聞'94 データ集の記事データと RWC テキストデータベース第 1 版 [5] の毎日新聞記事形態素解析差分データより、品詞情報付きの形態素データを抽出した。抽出したデータは、毎日新聞記事 94 年度の全記事のタイトルと本文の形態素解析である。形態素ののべ数は 30,782,848、異なり数は 229,305 である。内訳を図 2に示す。また、図 3が名詞の詳細区分である。「その他」以外は、抽出したデータで用いられている名称をそのまま記している。括弧内は合計に対する割合である。

品詞	数	品詞	数
名詞	152,604 (66.6)	接続詞	133 (0.06)
未知語	45,104 (19.7)	記号	110 (0.05)
動詞	27,153 (11.8)	助動詞	63 (0.03)
形容詞	2,104 (0.92)	感動詞	25 (0.01)
副詞	1,453 (0.63)	外字	10 (0.00)
接頭詞	252 (0.11)	空白	1 (0.00)
連体詞	150 (0.07)	その他	1 (0.00)
助詞	142 (0.06)	合計	229,305

図 2: RWC データに含まれる異なり形態素の内訳 (毎日新聞 94 年度全記事対象)

ii) 形態素データの辞書への取り込み

i) で抽出した形態素を本システムの辞書のフォーマットに変換した。形態素辞書で利用するタグセットは、RWC のタグセットを元に変更

名詞の詳細分類	数
固有名詞	71,474 (46.8)
名詞	37,515 (24.6)
数	34,384 (22.5)
サ変接続	5,218 (3.42)
(形容動詞語幹)	2,381 (1.56)
接尾	1,052 (0.69)
副詞可能	377 (0.25)
代名詞	98 (0.06)
その他	105 (0.07)
合計	152,604

図 3: 名詞の詳細区分

を加えたものである。取り込み後の形態素辞書の内訳は図4になる。変換する際に固有・普通名詞は取り除いた。図2の動詞、形容詞は語幹と活用語尾に分割したため、図2の動詞語幹、形容詞語幹との間には数に違いがある。括弧内は合計に対する割合である。

品詞	数	品詞	数
動詞語幹	6,652 (50.1)	連体詞	150 (1.13)
形容動詞語幹	2,381 (17.9)	助詞	142 (1.07)
		動詞	138 (1.04)
副詞	1,453 (11.0)	接続詞	133 (1.00)
接尾詞	1,052 (7.92)	記号	110 (0.83)
形容詞語幹	692 (5.21)	名詞	106 (0.80)
		感動詞	25 (0.19)
活用語尾	231 (1.74)	形容詞	2 (0.02)
合計	13,267		

図 4: 形態素辞書の内訳

2.2.2 連接規則、形態素の接続規則、係り受け規則

システム、形態素辞書の作成後、毎日新聞記事5件 (タイトル込で 78 文) を解析した。その結果に基づき、辞書とシステム作成時の係り受け規則を修正、追加した。そして、連接規則、形態素の接続規則の作成を人手で行なった。現時点での辞書の見出し語数は 12,189、連接規則数は 279、形態素の接続規則数は 59 である。また、係り受け規則はシステム内に組み込んであり、その数は約 50 である。

2.3 辞書、規則の構造

2.3.1 形態素辞書

形態素辞書の項目は「見出し語 (品詞, 活用型, 活用形)」からなる。それぞれの説明は以下の通りである。

- **見出し語:** 形態素の表記
- **品詞:** 形容詞、動詞等
- **活用型:** 品詞の詳細区分や、活用の区分
- **活用形:** 未然形、連用形、終止形等

本論文では、(品詞, 活用型, 活用形) の 3 つを合わせて「品詞情報」と呼ぶ。

形態素の例を以下に示す。

基づ (動詞語幹, 五段・カ行, ϕ) … (1)
く (活用語尾, 五段・カ行, 連体形) … (2)

"ϕ"は要素がないことを意味する記号である。

2.3.2 連接規則

連接規則の項目は「(P1, P2)」(P: 品詞情報) からなる。P1 を持つ形態素と、P2 を持つ形態素は連接可能である。

連接規則の例を以下に示す。

((動詞語幹, X, ϕ), (活用語尾, X, *)) … (3)

"ϕ"は要素がない、"*"は任意の要素が適用可能、"X"は規則中で同一の要素が適用可能であることを示す。

2.3.1 の 2 つの形態素 (1)(2) は、「五段・カ行」という活用型を持つ動詞語幹と活用語尾 (活用形は連体形) であり、(3) 規則により連接可能である。

2.3.3 形態素の接続規則

形態素の接続規則は、接続する複数の形態素の品詞情報と、接続後の品詞情報を記述する。形態素の接続規則の項目は「(P1 + … + Pn → P')」(P: 品詞情報) からなる。それぞれの説明は以下の通りである。

- **P1 + … + Pn:** 連続する複数の品詞情報であり、「接続部分」と呼ぶ。
- **P':** 接続した結果となる語の品詞情報であり、「置換部分」と呼ぶ。

形態素の接続規則の例を以下に示す。

((動詞語幹, X, ϕ)+(活用語尾, X, Y)→(動詞, X, Y))…(4)

"ϕ"、"*"、"X"の意味は 2.3.2 と同じであり、"Y"の意味は "X"と同じである。

形態素解析の結果として 2.3.1 の 2 つの形態素 (1)(2) が現れるとき、規則 (4) によって 1 つの動詞「基づく (動詞, 五段・カ行, 連体形)」に置換される。

2.3.4 係り受け規則

係り受け規則は語の品詞情報を利用して、語の係り先を決める規則である。以下、係り受け規則の例を示す。

- 名詞 ⇒ 直後の語に係る。 … (5)
- 格助詞 ⇒ 直後の述語に係る。 … (6)

- 活用形が連体形の語 ⇒ 直後の名詞に係る。… (7)
- 活用形が連用形の語
  ⇒ 直後の連体形以外の述語に係る。 … (8)

係り受け規則 (5)～(8) を利用して係り受け解析を行なった例を、図5に示す。図5では矢印が係り元の語と係り先の語を結び、下の数字が利用する係り受け規則を表す。

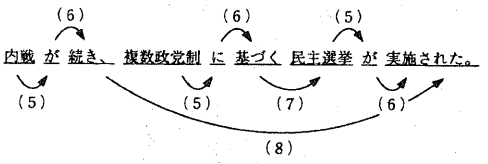

図 5: 係り受け例

2.4 処理アルゴリズム

2.4.1 形態素解析

形態素解析は形態素辞書、連接規則を利用して、図6のアルゴリズムに従って行う。

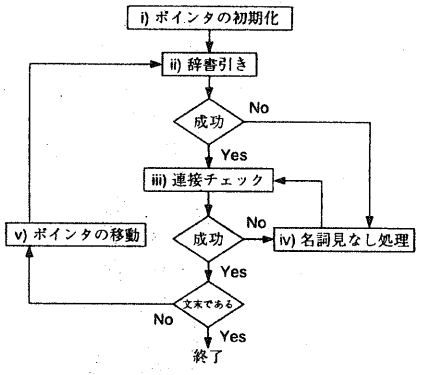

図 6: 形態素解析アルゴリズム

i) **ポインタの初期化** … 処理位置を指すポインタを、文の先頭の文字にセットする。

ii) **辞書引き** … ポインタが指す位置から始まる文字列について、最長一致によって辞書 (図1の形態素辞書を指す) 引きを行なう。

iii) **連接チェック** … 辞書引き、または名詞と見なした形態素と1つ前に確定した形態素との連接可能性を、連接規則を用いてチェックする。

iv) **名詞見なし処理** … ポインタが指す1文字を名詞とし、1つ前に確定した形態素が名詞で文字種 (カタカナ、アルファベット、記号、それ以外) が同じであれば2つをつなげる。

v) **ポインタの移動**… ポインタを文末に向かって、1つ隣の文字に移す。

連接チェック時に接続可能な形態素が複数存在する場合、最初に辞書引きされた形態素をその時点の形態素とし、解析を進める。形態素は見出し語の文字コード順にソートされている。以下の処理では、解析の曖昧性は保持しない。

2.4.2 形態素の接続

形態素の接続規則を利用して、図7のアルゴリズムに従って行う。

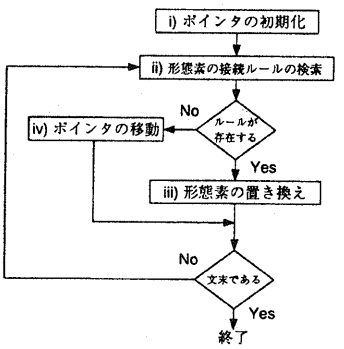

図 7: 形態素の接続アルゴリズム

i) **ポインタの初期化** … 処理位置を指すポインタを、文の先頭の形態素にセットする。

ii) **形態素の接続規則の検索** … 形態素の接続規則の中から、規則の接続部分が、ポインタが指す形態素から連続する n 個 (n は最大) に適応できるものを探す。

iii) **形態素の置換** … ii) で検索した規則の一致部分に当たる形態素を、置換部分の形態素で置き換え、置換した形態素をポインタで指す。

iv) **ポインタの移動** … ポインタを文末に向かって、1つ隣の形態素に移す。

形態素の接続規則の検索時に、適応可能な規則が複数存在する場合、最初に検索した規則を適応する。形態素の接続規則は文字コード順にソートされている。

2.4.3 構文解析

構文解析は係り受け規則を利用して、図8のアルゴリズムに従って行う。

i) **ポインタの初期化** … 処理位置を指すポインタを、文の最後の語にセットする。

ii) **係り受け先の決定** … 係り受け規則を利用して、語の係り受け先を決定する。

iii) **ポインタの移動** … ポインタを文頭に向かって、1つ隣の語に移す。

語の係り受け先の決定に利用する係り受け規則は、語の見出し語、品詞情報を利用して、システム内で1つに絞られる。

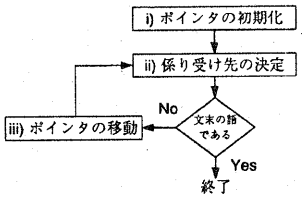

図 8: 構文解析アルゴリズム

3 JEMONI の評価

3.1 名詞連続部分の推定精度

JEMONI を情報検索で利用する際には、索引語 (index) として利用する名詞部分を正しく抽出することが重要である。そこで名詞連続部分の推定精度に基づいて評価を行なう。

3.1.1 評価方法

RWC テキストデータベースの新聞記事 3,000 件に対する毎日新聞形態素解析差分データ (以下、RWC データ) と、京都大学テキストコーパス[1][3] の新聞記事約 20,000 文に対する形態素解析結果 (以下、京大コーパス) を正解データとする。評価尺度には再現率、適合率を用いる。

$$\text{再現率} = \frac{\text{正しく推定できた名詞連続の総数}}{\text{正解データ中の名詞連続の総数}}$$

$$\text{適合率} = \frac{\text{正しく推定できた名詞連続の総数}}{\text{出力した名詞連続の総数}}$$

3.1.2 評価結果および考察

単一の名詞についても、3.1.1 と同様の式を利用して比較した[2]。名詞連続推定の評価結果との差異を明らかにし、JEMONI の特性をより明らかにするためである。結果を表 1に示す。

RWC データに対する結果では、茶筌、JUMAN の解析結果中の総正解名詞数は JUMAN の方が 300 弱多い。JUMAN は辞書に持たない語を未定義語とし、名詞としてはカウントせず、正解から漏れている。一方、茶筌は同じく辞書に持たない語を名詞とし、正解としている。

[1]Version 2.0。毎日新聞記事 95 年度の 1 月 1 日から 1 月 10 日までの全記事 (文数: 10,723、1 文あたりの平均文字数: 44.6)、1 月 11 日から 6 月 30 日までの社説記事 (文数: 9,233、1 文あたりの平均文字数: 41.5) を対象とする。

[2]QJP の解析結果では、形態素の品詞で、名詞と後方一致するもの (例. 数名詞、形式名詞、時副詞名詞) は名詞として数えた。

表 1: 名詞推定の評価

評価対象	再現率 (%)	適合率 (%)
茶筌 1.51	$\frac{239,935}{377,464}$ (63.6)	$\frac{239,935}{359,202}$ (66.8)
JUMAN 3.5	$\frac{240,213}{377,464}$ (63.6)	$\frac{240,213}{302,953}$ (79.3)
QJP 1.50	$\frac{116,987}{377,464}$ (31.0)	$\frac{116,987}{231,845}$ (50.5)
JEMONI	$\frac{159,046}{377,464}$ (42.1)	$\frac{159,046}{287,753}$ (55.3)

(a) RWC データ

評価対象	再現率 (%)	適合率 (%)
茶筌 1.51	$\frac{145,275}{178,487}$ (81.4)	$\frac{145,275}{188,661}$ (77.0)
JUMAN 3.5	$\frac{163,043}{178,487}$ (91.3)	$\frac{163,043}{167,390}$ (97.4)
QJP 1.50	$\frac{92,625}{178,487}$ (51.9)	$\frac{92,625}{138,445}$ (66.9)
JEMONI	$\frac{93,659}{178,487}$ (52.5)	$\frac{93,659}{163,873}$ (57.2)

(b) 京大コーパス

JUMAN の未定義語を名詞としてカウントすると、RWC コーパスに対して再現率は $\frac{259,584}{377,464}$ (68.8%)、適合率は $\frac{259,584}{330,097}$ (78.6%)、京大コーパスに対して再現率は $\frac{174,042}{178,487}$ (97.5%)、適合率は $\frac{174,042}{178,768}$ (97.4%) になる。

また、茶筌は複数の文字からなる数詞を、1 文字の名詞として切り分けている。

例.

JUMAN				茶筌	
68	名詞	数詞	→	6	名詞 数詞
				8	名詞 数詞

茶筌の出力名詞数が他と比べて多いのは、それが理由である。QJP、JEMONI の結果は、茶筌、JUMAN よりも低い。これは複合名詞を切り分けている、いないの差である。どの評価対象システムの結果も、RWC データに対するより、京大コーパスに対する方が精度が高い。この原因は 2 つ考えられる。1 つは、RWC データでは接尾を名詞の下位区分として取り扱い、京大コーパスに対し各評価対象のシステムでは、接尾に対応する品詞[3]は名詞や動詞と同列に取り扱っているからである。RWC データに対する評価では、各評価対象システムの接尾に対応する形態素が正解とカウントされないので、京大コーパスに対するよりも再現率、適合率を下げる結果となっている。もう 1 つは、京大コーパスでは

[3]京大コーパス、茶筌、JUMAN では接尾辞、QJP では品詞情報で「=辞」と後方一致するもの (例. 名詞=尾、時副詞数名詞=尾)。JEMONI では接尾詞。

テキストコーパスとして不適当な文中の箇所、文、記事を削除しているのに対して[4]、RWCデータでは取り除いていないことである。それらがRWCデータに対する解析精度を下げる結果となっている。

名詞連続推定の評価結果を表2に示す。これは評価対象の解析結果、正解データ中の隣り合う名詞を接続し、1つの名詞として扱った場合の値である。

表 2: 名詞連続推定の評価

評価対象	再現率 (%)	適合率 (%)
茶筌 1.51	$\frac{174,987}{242,549}$ (72.1)	$\frac{174,987}{235,811}$ (74.2)
JUMAN 3.5	$\frac{154,054}{242,549}$ (63.5)	$\frac{154,054}{224,188}$ (68.7)
QJP 1.50	$\frac{109,988}{242,549}$ (45.3)	$\frac{109,988}{196,199}$ (56.1)
JEMONI	$\frac{179,353}{242,549}$ (73.9)	$\frac{179,353}{260,063}$ (69.0)

(a) RWCデータ

評価対象	再現率 (%)	適合率 (%)
茶筌 1.51	$\frac{113,867}{135,026}$ (84.3)	$\frac{113,867}{134,276}$ (84.8)
JUMAN 3.5	$\frac{122,056}{135,026}$ (90.4)	$\frac{122,056}{128,266}$ (95.2)
QJP 1.50	$\frac{100,429}{135,026}$ (74.4)	$\frac{100,429}{121,346}$ (82.8)
JEMONI	$\frac{107,615}{135,026}$ (79.7)	$\frac{107,615}{152,251}$ (70.7)

(b) 京大コーパス

茶筌、QJP、JEMONIの再現率、適合率は、RWCデータ、京大コーパスを正解データとした両方の場合で、表1よりも高くなっている。これは名詞を接続することで、複合名詞の分割誤りを吸収するからである。例えば、茶筌、QJP、JEMONIでは、正解データ中の「東京ドーム」を「東京」と「ドーム」に、「光ファイバー」を「光」と「ファイバー」に分割してしまう。それらは、表1の名詞推定の評価では誤りとしてカウントされるが、表2の名詞連続推定の評価では正解とカウントされる。

一方、JUMANの再現率、適合率は、表1よりも表2の方が低い。これはJUMANでも、上記の茶筌、QJP、JEMONIのように、複合名詞の分割誤りは吸収するが、辞書に見出し語を持たない語を未定義語とすることによって、精度が下がるからである。例えば、JUMANで「リベラル (未定義語) 新党 (名詞)」と解析したときに、名詞の推定では、正解が「リベラル (名詞) 新党 (名詞)」となり1つ正解するが、名詞連続の推定では、正解は「リベラル新党 (名詞)」となり、JUMANの結果は誤りとなる。また、JUMANの未定義語を名詞としてカウントすると、RWCコーパスに対して再現率は$\frac{171,968}{242,549}$ (70.9%)、適合率は$\frac{171,968}{237,493}$ (72.4%)、京大コーパスに対して再現率は$\frac{132,950}{135,026}$ (98.5%)、適合率は$\frac{132,950}{134,914}$ (98.5%) になる。これは表1のJUMANの結果に対して、未定義語を名詞としてカウントして出した結果よりも、RWCコーパスに対して再現率が約2%高く、京大コーパスに対しても再現率、適合率が約1%高い。

JEMONIは名詞辞書を持たなくても、名詞連続を他のツールと同程度に推定できる。

3.2 係り受け先の解析精度

3.2.1 評価方法

京大コーパスを正解データとする。はじめに京大コーパスの対象とする約20,000文の新聞記事に対して、KNP、QJP、JEMONIで解析を行なう。次に京大コーパスと各構文解析結果を共通のフォーマットに変換する。そしてフォーマット変換後、各構文解析結果と正解データを比較する。

比較は文節単位と文単位でそれぞれ行なった。文節単位の比較での評価尺度は再現率、適合率である。

$$\text{再現率} = \frac{\text{正しく係り受け先を特定できた文節の総数}}{\text{正解データ中の文節の総数}}$$

$$\text{適合率} = \frac{\text{正しく係り受け先を特定できた文節の総数}}{\text{出力した文節の総数}}$$

文単位での評価値は正解率 (accuracy) である。

$$\text{正解率} = \frac{\text{全ての係り受け先を正しく特定できた文数}}{\text{京大コーパス中の文数}}$$

KNP、QJPは文節単位で係り受け先を解析する。JEMONIは2.2.2で述べた5件の新聞記事を解析した範囲で、形態素の接続規則に文節をつくる規則を追加した。

3.2.2 評価結果および考察

それぞれの評価結果を表3に示す。

文節単位、文単位ともに、QJP、JEMONIの解析精度は、KNPの解析精度よりも低い。QJPは形態素解析の精度がKNPの形態素解析 (JUMAN) の精度よりも低いこと、会話文 (「」内の文) 外の文節が会話文内の文節に係る場合があること、並列構

[4]記事のタイトル、記事中の小見出し (例. ◇国民自治に乗り出そう)、漢字名に対する読み、"()"内にある補足等やアンケート記事、広告記事等。

表 3: 京大コーパスと各解析システムとの比較結果

解析システム	再現率 (%)	適合率 (%)
KNP 2.0b6	$\frac{174,955}{192,366}$ (90.9)	$\frac{174,955}{192,551}$ (90.9)
QJP 1.50	$\frac{127,242}{192,366}$ (66.1)	$\frac{127,242}{196,822}$ (64.6)
JEMONI	$\frac{88,075}{192,366}$ (45.8)	$\frac{88,075}{221,806}$ (39.7)

(a) 文節単位

解析システム	正解率 (%)
KNP 2.0b6	$\frac{11,202}{19,956}$ (56.1)
QJP 1.50	$\frac{4,366}{19,956}$ (21.9)
JEMONI	$\frac{1,586}{19,956}$ (7.95)

(b) 文単位

造解析の精度が低いことが、解析精度を下げる原因となっている。JEMONI は形態素解析の精度が低いこと、係り受けルールの不備・不足、並列構造の解析をしていないことが、解析精度を下げる原因となっている。

4 voting 法

本論文では、評価の 1 手法である voting 法を提案する。voting 法とは、複数システムの共通解 (以下、voting 結果) を正解データとする評価方式である。情報検索の評価、解析システムの評価等、競合するシステムが複数存在する場合は、評価対象を限ることなく利用できる。また共通解を自動的に正解とするため、評価用コーパスを自動的に生成するという特長を持つ[5]。

以下、構文解析の評価を取り上げて voting 法の妥当性を検証し、voting 法を利用した評価例として、JEMONI の係り受け先の解析精度を評価した。それぞれ 4.1、4.2 に示す。

4.1 voting 法の妥当性

KNP と QJP の構文解析結果の共通解を voting 結果とし、京大コーパスと比較した。voting 結果の精度が高く、解析対象となるテキストの性質をよく捉えていれば、信頼できる正解データであり、voting 結果を利用した評価は意義がある。

voting 結果の作成方法は以下の通りである

[5]情報検索のテストコレクション作成に用いる「pooling 方式」とは異なり、本手法は人間が介在しない。

i) KNP、QJP で、京大コーパスが対象とする約 20,000 文を構文解析する。

ii) i) の両構文解析結果を共通のフォーマットに変換する。共通フォーマットは係り元の文節と係り受け先の文節の対からなり、各文節の先頭には文中での位置が付与されている。

例.

位置:係り元の文節	位置:係り先の文節
0:あらかじめ	10:発表された
10:発表された	20:十人を
20:十人を	38:投票する
26:読者が	38:投票する

iii) ii) それぞれの結果から、文節単位の共通部分、文単位の共通部分を抽出する。

以上の方法により、KNP、QJP の解析結果の積から作成されたものが、ここでの voting 結果となる。voting 結果を京大コーパスと比較し、文節単位、文単位でのカバレッジ (coverage) と正解率 (accuracy) を、以下の式から算出した。

$$\text{カバレッジ} = \frac{\textit{voting}\text{ 結果中の総文節 (文) 数}}{\text{京大コーパス中の総文節 (文) 数}}$$

$$\text{正解率} = \frac{\text{分母のうちで正解となる文節 (文) 数}}{\textit{voting}\text{ 結果中の総文節 (文) 数}}$$

voting 法の評価結果を表 4に示す。

表 4: voting 法の評価

	文節単位	文単位
カバレッジ (%)	$\frac{127,549}{192,366}$ (66.3)	$\frac{4,497}{19,956}$ (22.5)
正解率 (%)	$\frac{122,428}{127,549}$ (96.0)	$\frac{3,853}{4,497}$ (85.7)

カバレッジは文節単位、文単位でともに低い。しかし、人間が作成した正解データに近い精度 (正解率) を持つテストデータが作成できた。

4.2 voting 法による評価例: JEMONI の係り受け先の解析評価

4.1 節で作成した voting 結果を利用し、JEMONI の係り受け先の解析評価を行なった。評価値には文節単位、文単位での正解率 (accuracy) を用いている。算出式は以下の通りである。

$$\text{正解率} = \frac{\text{正解となる文節 (文) の総数}}{\textit{voting}\text{ 結果中の総文節 (文) 数}}$$

評価結果を表 5に示す。

結果を単純に比較することはできないが、文節単位、文単位での正解率は、京大コーパスとの評価の値よりも高くなっている。これは voting 結果に解析の易しい文が集まったことが理由である。

表 5: voting 法による JEMONI の評価

	文節単位	文単位
正解率 (%)	$\frac{73,232}{127,549}$ (57.4)	$\frac{1,188}{4,497}$ (26.4)

5 おわりに

本論文では、我々が開発した形態素・構文解析システム JEMONI を紹介し、既存の形態素・構文解析システムと名詞連続の推定精度、係り受け先の解析精度の観点から比較評価を行なった。そして、複数システムの共通解を正解データとする評価の一手法、voting 法を提案した。

JEMONI の解析精度は他の公開システムと較べると低く、改良の余地がある。また voting 法により、カバレッジは低いが、人間が作成した正解データに近い精度 (正解率) を持つテストデータが作成できた。

謝辞

JUMAN、茶筌、KNP の精度向上に日々努力を続けている皆様と、QJP の開発者であるリコーの亀田雅之氏に感謝の意を表します。

参考文献

[1] 黒橋禎夫, 長尾真: 日本語形態素解析システム JUMAN version 3.5, 京都大学大学院工学研究科 (1997)

[2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明: 日本語形態素解析システム『茶筌』 version 1.5 使用説明書, Information Science Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学 (1997)

[3] 黒橋禎夫: 日本語構文解析システム KNP version 2.0 b6 使用説明書, 京都大学大学院 情報学研究科 (1998)

[4] 亀田雅之: 軽量・高速な日本語解析ツール『簡易日本語解析系 Q_JP』, 言語処理学会 第 1 回年次大会, pp. 349-351 (1995)

[5] 新情報処理開発機構: RWC テキストデータベース報告書 (1996)

[6] 黒橋禎夫, 齊藤由衣子, 坂口昌子: コーパス作成の作業基準 version 1.6, 京都大学 (1998)

付録

JEMONI の処理結果例を図 9 ～11に示す。

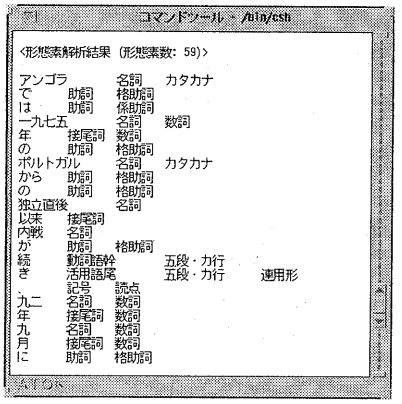

図 9: 形態素解析結果例

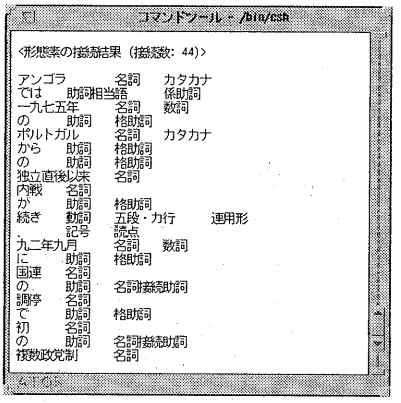

図 10: 形態素の接続結果例

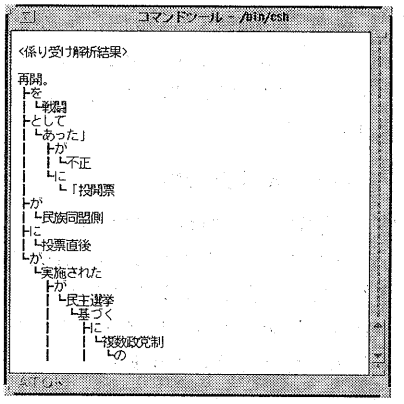

図 11: 構文解析結果例