TR-IT-0192

TDMT 用形態素解析プログラム (日本語 / 英語 / 韓国語) — 性能評価報告 —

Performance evaluation of the morphological analyzer for TDMT (Japanese/English/Korean)

山本 和英　　隅田 英一郎

Kazuhide YAMAMOTO and Eiichiro SUMITA

1996 年 9 月

概要

変換主導型機械翻訳 (TDMT) 用の形態素解析プログラムの性能について報告する。本システムは高い頑健性を目指して品詞と単語の混合 n-gram モデルを採用しており、現 TDMT システムが翻訳対象としている日本語、英語、韓国語をほぼ同一の機構によって処理している。評価結果の概要は以下の通り。日本語は公開されている他の高精度の形態素解析プログラムと比較してほぼ同等の精度であり、かつ他とは異なる話し言葉固有の表現も処理できることが確認でき、解析時間も問題なかった。英語は解析能力、解析時間共に全く問題なく、韓国語はこれら二言語に準ずる性能を示してはいるがまだ向上の余地があることが明らかになった。

エイ・ティ・アール音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

もくじ

第 1 章

形態素解析概要

1.1 特徴

従来の形態素解析システムと比較した場合の本形態素解析システムの特徴は以下の通りである。

1. 話し言葉を対象としている。
2. 言語モデルをタグ付きコーパスから学習している。
3. 未知データに対しても適応できる高い頑健性を持ったシステムの実現のために、品詞と単語の混合 n-gram を言語モデルとして採用している。
4. 日本語、英語、韓国語をほぼ同一の手法によって処理している。

1.2 アルゴリズム

本形態素解析は、品詞と単語の混合 n-gram に基づいて入力文の尤度を計算し、その尤度が高い候補を出力する。

$$\text{累積尤度} = \sum_{\text{形態素列}} \text{各形態素選択の尤度}$$

$$\text{形態素選択の尤度} = \log \frac{\text{選択形態素を}n\text{語目とするn-gram頻度}}{\text{直前の}(n-1)\text{-gram頻度}}$$

言語モデル 助詞や助動詞などの機能語に対しては、表記を含む各項目のいずれかが異なるものは別単位として、名詞や動詞などの内容語に対しては、品詞と活用情報のいずれかが異なるもののみを別単位として n-gram を作成する。

未知語処理 現在はまだ特別な処理は行なっていない。辞書登録されていない語の出現には、未知語として出力を行なうか、形態素解析出力に失敗するかを選択できる。

ビーム幅 全数サーチではなく、ビームサーチによって処理を行なっている。ビーム幅で設定された尤度の低い中間候補を切り捨てる。

第 2 章

性能評価

2.1 評価尺度

今回の形態素解析の評価尺度として、以下の項目を用いた。

- 処理速度 (平均速度、最悪速度。単位 sec.)
- 精度
  - 再現率 (recall) : $Rcl.$

$$Rcl. = \frac{M_{match}}{M_{tagged}}$$

  - 適合率 (precision) : $Pcn.$

$$Pcn. = \frac{M_{match}}{M_{output}}$$

  - 文正解率 (sentence accuracy) : $S.Ary.$

$$S.Ary. = \frac{S_{match}}{S}$$

ただし、

M_{tagged}:	正解形態素数
M_{output}:	出力形態素数
M_{match}:	正解と出力で一致する形態素数
S:	正解文数 (= 出力文数)
S_{match}:	正解と出力で一致する文数

なお、計算上の注意として、以下の 2 点がある。

1. (韓国語においては) 出力形態素によっては、M_{match} に複数の数え方が存在する可能性がある。この場合は、それらのうち最も高い値を M_{match} として計算した。

2. 形態素解析に失敗し、何も出力しなかった文 (以下、無出力文と呼ぶ) は、正解の形態素数と同数の形態素出力があり、かつすべて正解と異なる出力であったとみなして計算した。

今回の評価では、出力候補のうち最も評価値の高い出力 (以下、 Top-1 と呼ぶ) だけでなく、評価値が上位 n 位以内の出力 (以下、 Top-n,n=2,3,$\cdots$) も考慮した。その場合、 Top-n のうち最も適合率の高い候補を Top-n の代表とし、その代表と正解から計算される再現率、適合率、文正解率を Top-n の再現率、適合率、文正解率とした。

2.2 評価方法

今回の評価では、日本語、英語、韓国語それぞれの形態素解析システムの比較を行なうため各言語共通の評価と、システムに関係する各要素の影響を見るための個別の評価を行なった。

3 言語共通で行なった評価実験の仕様は以下の通りである。

- 言語モデルは混合 bi-gram を採用。
- テキストは ATR 旅行会話コーパスから選択。
- 使用テキストは 4000 文とする。これは言語別に独自に選択する。文 ID は一致させないという条件で選択したので、同一の ID を持つ文を複数回選択することはないが、偶然内容が同一である可能性はある。
- 実験は選択したテキストを対象にした 10 分割の cross validation によって行なう。つまり言語モデルのための学習文数 3600 文、テスト文数 400 文の実験を、組合せを変えて 10 回行ない、その平均を実験結果とする。
- ビーム幅は 30。ただし、 Top-100 の実験のビーム幅は 100。

この他に、システム要素の影響を見るために行なった実験は以下の通りである。これらの実験は、それぞれ実験可能な 1 言語のみで行なった。

- 学習量との関係
- 言語モデルとの関係
- ビーム幅との関係

2.3 評価結果

2.3.1 概要

各種実験により得られた結果をまとめると以下の通り。

処理速度 平均で日本語 0.2 秒、英語 0.015 秒、韓国語 0.3 秒。
韓国語の最悪時間が遅いが、平均はほぼ問題なし。

Top-1 精度

単語適合率 (Pcn.) 日本語 98.0%、英語 97.5%、韓国語 93.5%。
韓国語は日英両言語の精度に比べ改善の余地は大きい。

文正解率 (S.Ary.) 日本語 85.8%、英語 88.8%、韓国語 72.0%。
ただしテストセットの文長が言語間でばらつきがあるため相互比較不可。

Top-n 精度 n 増大に伴い向上し、Top-100 でも収束していない。
特に文正解率にこの傾向あり。

学習量との関係 32000 文で、単語適合率が 98.9%、文正解率で 90%。
対数的にではあるが、まだしばらくは学習量増加により精度向上が見込める。

2.3.2 詳細

共通実験諸元

表 2.1: 共通実験諸元

項目	日本語	英語	韓国語
品詞数	32	35	33
付与要素数 (品詞+属性)	77	67	64
辞書語彙数 (基本形のみ)	8392	2623	2428
辞書語彙数 (活用形含む)	9729	3349	–
bi-gram の Perplexity	2.937	8.507	6.631
学習文数	3600 文 / 回		
のべ形態素数 (平均)	41339	23171	32305
異なり形態素数 (平均)	3763	2405	2134
テスト文数	400 文 / 回		
のべ形態素数 (平均)	4593	2575	3589
異なり形態素数 (平均)	1075	737	763
文平均形態素数	11.48	6.43	8.97
使用計算機	Sun SPARCstation 10		
設定ビーム幅	30(Top-100 の時は 100)		

- 本形態素解析では、分割した形態素に品詞だけでなく、活用形など[1]の属性も同時に付与し、これが誤った場合も形態素解析の誤りとしている。これらの属性も含めた場合の数は、形態素解析の精度と関係があることから、上の表では「付与要素数」として明記した。

- 田代らの論文 [田代 96] では、ATR 体系での品詞数 32、異なり語数 2908、また EDR 体系での品詞数 15、異なり語数 2963 とある。また、Nagata[Nag94] の論文では、ATR 体系を使用し、語数 6580 とある。[丸山 94] では、品詞数が自立語 50、活用語尾を含む付属語 71 で、合計 121 要素である。

- 日本語、英語の場合、活用も含んだ形で実際の辞書には登録されている。このため、本来の語彙数と活用も含めた語彙数の両者を併記した。なお、韓国語については、活用を含まない形で辞書に登録しているため、登録語彙数は 1 種類である。

- Perplexity は日本語で「複雑度」と呼び [中川 88]、情報理論的な意味での平均分岐数を意味する。上記の結果は、本形態素解析に使用した混合 bi-gram が英語、韓国語、日本語の順で複雑であることを意味する。

[1] 日本語は活用形 (連用形など)、英語は意味活用 (過去など) 及び一致活用 (三人称単数など)、韓国語は語彙情報。

処理速度

表 2.2: 共通実験の処理速度 (単位 sec.)

	日本語		英語		韓国語	
Top-n	平均	最悪	平均	最悪	平均	最悪
1	.206	1.7	.0152	0.15	.299	8.65
5	.207	1.65	.0161	0.167	.299	8.47
10	.206	1.683	.0173	0.167	.298	8.40
100	.558	3.817	.0332	0.566	.324	8.82

- 全言語において Top-100 の速度が遅いのは、ビーム幅を 100 に拡大しているためである。
- 英語の速度は全く問題なし。

処理精度

表 2.3: 共通実験の精度評価

	日本語			英語			韓国語		
Top-n	Rcl.	Pcn.	S.Ary.	Rcl.	Pcn.	S.Ary.	Rcl.	Pcn.	S.Ary.
1	.9826	.9805	.858	.9778	.9749	.8878	.938	.935	.720
5	.9949	.9943	.954	.9956	.9948	.9743	.975	.965	.883
10	.9962	.9959	.966	.9979	.9976	.9880	.979	.969	.903
100	.9986	.9984	.986	.9997	.9995	.9980	.989	.984	.948

- 多くの場合において、英語、日本語、韓国語の順で精度がよい。
- ただし、Top-1 での再現率、適合率は日本語と英語が逆転している (図 2.1を参照)。この理由は現在のところ不明。
- 文正解率は、英語、日本語、韓国語の順だが、テストセットの文長が言語によってばらつきがあるため相互に比較することはできない。

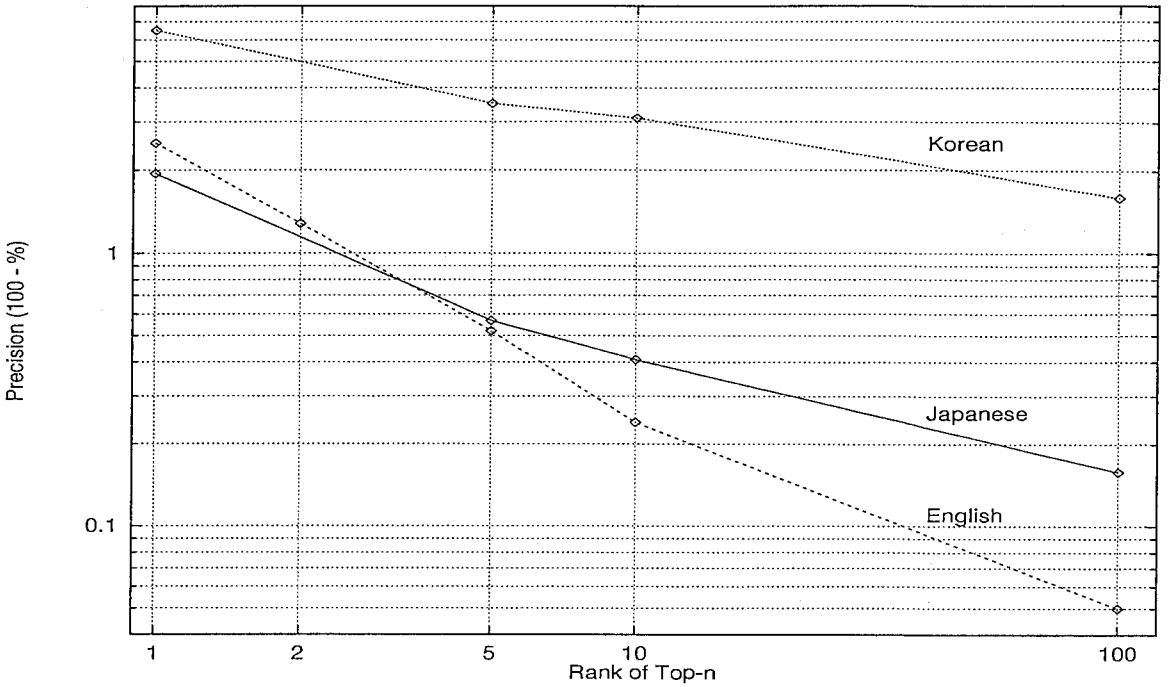

図 2.1: 日英韓三ヶ国語の適合率 (Pcn.) の比較

無出力文数

表 2.4: 無出力文数

Top-n	日本語	英語	韓国語
1	0	0	18
5	0	0	18
10	0	0	18
100	0	0	15

- 日本語、英語で無出力文がないのは出現語がすべて既知のため。このため、現在未知語処理は行なっていない。
- 韓国語で出力に失敗するのはコーパスの誤り、Tag 付けの誤りなどが原因と予想される。

学習量との関係

日本語を対象にして、学習量との関係を見るためにテキスト量を変化させた実験を行なった。共通実験と異なる点は使用した文数のみであり、10 分割の cross validation を行なっている点は同様である。

表 2.5: 学習量を変化させた場合の精度 (日本語)

	Top-1			Top-5		
文数	Rcl.	Pcn.	S.Ary.	Rcl.	Pcn.	S.Ary.
1000	.9726	.9682	.790	.9914	.9901	.9219
4000	.9826	.9805	.858	.9949	.9943	.9540
10000	.9845	.9834	.873	.9971	.9967	.9717
32561	.9887	.9888	.903	.9985	.9984	.9858

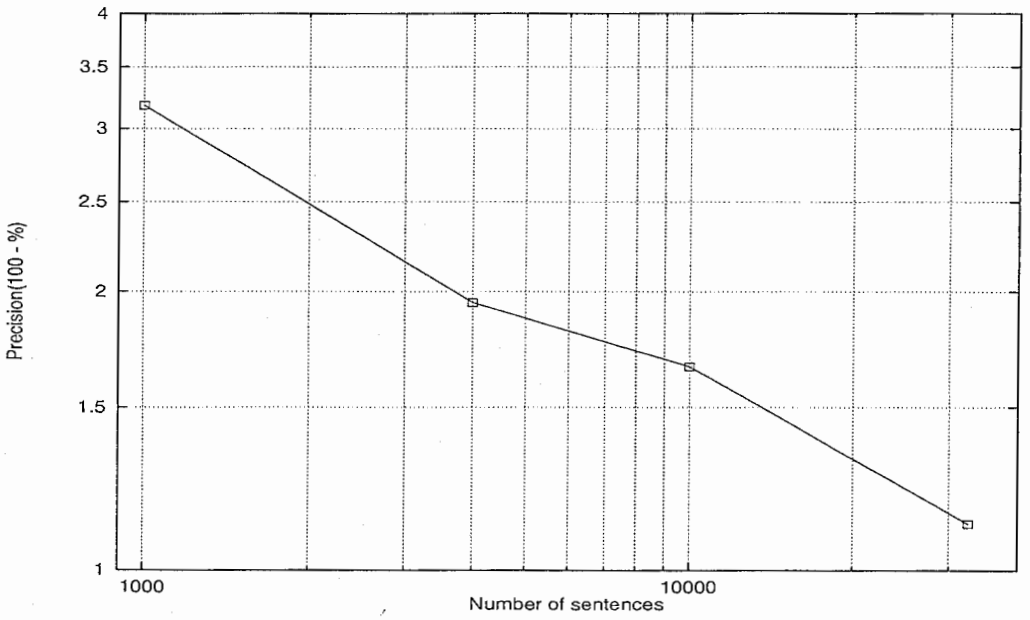

図 2.2: 学習量との関係 (日本語)

- 日本語において、学習量増加に伴い誤りも減少している。その割合はほぼ、学習量 10 倍につき、誤りが $\frac{1}{2}$。
- この傾向はしばらくの間は続くと予想される。

言語モデルとの関係

英語を対象にして、言語モデルとの関係を見るために、言語モデルに tri-gram を使用した実験を行なった。共通実験と異なる点は使用した言語モデルのみである。

表 2.6: tri-gram モデルの精度 (英語)

Top-n	Rcl.	Pcn.	S.Ary.
1	.975	.969	.878
2	.989	.987	.945
5	.996	.995	.975
10	.9981	.9979	.9888
100	.99989	.99989	.9993

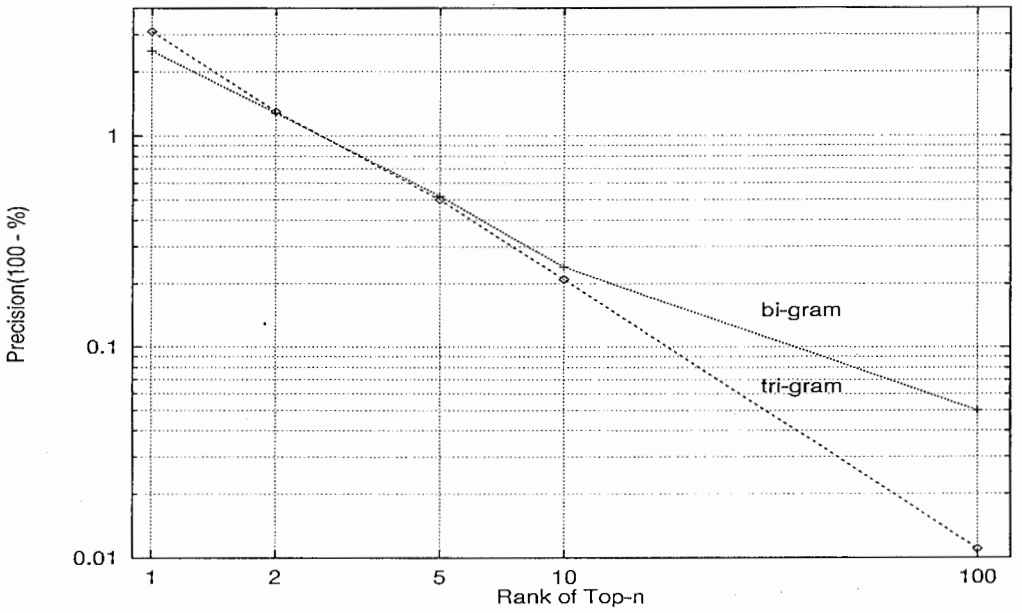

図 2.3: bi-gram と tri-gram の比較 (英語)

- 英語において、現在の品詞体系、コーパス量では両者の差は明確でない。
- Top-1 から Top-10 あたりまではほぼ同一の精度を出しているが、Top-100 では適合率、文正解率ともに tri-gram のほうが良好な結果になっている。

ビーム幅との関係

韓国語を対象にして、ビーム幅との関係を見るために、ビーム幅 100 を使用した実験を行なった。共通実験と異なる点はビーム幅のみ (共通実験ではビーム幅 30) である。

表 2.7: ビーム幅が 100 の場合の精度 (韓国語)

Top-n	Rcl.	Pcn.	S.Ary.
1	.941	.942	.727
5	.978	.972	.896
10	.984	.978	.922
100	.989	.984	.948

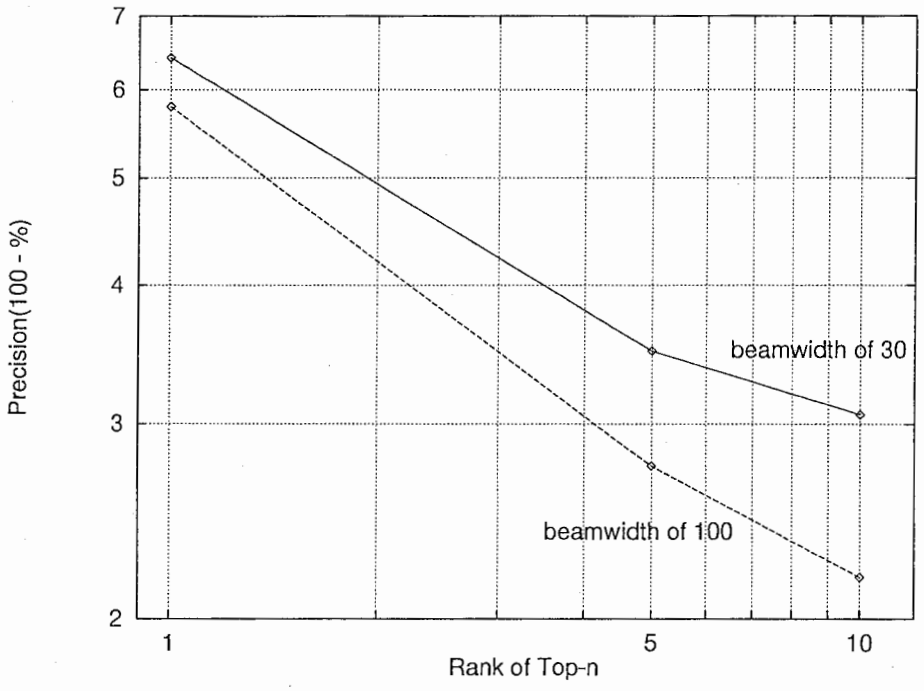

図 2.4: ビーム幅と適合率 (Pcn.) の関係 (韓国語)

- 韓国語において、ビーム幅の増加によって精度も向上するが、これによる寄与率は低い。
- 時間的な制約もあることも考慮に入れると、少なくとも韓国語に関してはビーム幅 30 程度で十分である。

第 3 章

日本語形態素解析の分析

3.1 正解の分析

本形態素解析システムにおいて正しく出力する文のうち、話し言葉に特徴的な文例を以下に示す。

いわゆる「ら抜き言葉」

- 辛いものは食べれますか。
- 誰でも見れるのですか
- このシールは何にでも付けれます。

感動詞、あいさつ、話し言葉特有の接続詞

- あっここでいいです。
- いや、そうですか。
- お、それはどうもきがつきませんでした。
- いえ、とんでもございません。
- どれどれ、ここにありました。
- うーん、それは、だから、よくわかんないんですけど。
- やれやれ、もうほんとに気分が悪くって、動けないんですよ。
- そりゃやっぱり市場へは行ってほしいです。
- じゃ、いってらっしゃい。
- てことはどっちがどんなに安いというか金額はどれくらいなんですか。
- ごめんくださいませ。

話し言葉特有の助詞、判定詞

- 京都だけじゃちょっともの足りないんですが。
- この大雪じゃあ仕方ないですよね。
- 十月十日っていやあ...
- 日本じゃ体育の日なんですけどねえ、
- うーん、それは、だから、よくわかんないんですけど。
- なんだか気持ち悪くってしょうがないんですよ。
- フイルムったって、いくらにもならないじゃないですか。

その他話し言葉特有の単語

- しょうがないですね。(形容詞「しょうがない」)
- よーく聞いてくださいね。(形容詞「よーく」)
- それはどんなもんですか。(形式名詞「もん」)
- 料理を作っていただけるようなとこはございませんか。(形式名詞「とこ」)
- こっから行こうとすれば...(代名詞「こっ」)
- フランスっぽいですね。(準体助動詞「っぽい」)
- お急ぎでらっしゃいますか。(補助動詞「らっしゃる」)
- そのあと銀行にいらしても十分間に合います。(本動詞「いらす」)

3.2 誤りの分析

3.2.1 誤りの分類

日本語形態素解析において誤っている部分の多くは、ひらがなで表記されている部分である。ここではこれらの誤りを分類して、問題解決の可能性を議論する。

1. 分割誤り
   「に (格助詞) なる (本動詞)」と「になる (補助動詞)」、
   「週 (普通名詞) 間 (普通名詞)」と「週間 (普通名詞)」など

- 数は少ない。
  本質的に区別が難しいものも多いが、人手で付与された Tag または仕様の一部に問題がある可能性もあり、さらに検討を進める必要がある。

2. 品詞付与誤り
   誤りの多くを占める。これはさらに二分できる。

   (a) 助詞細分類間の誤り
   「と」(格助詞と並立助詞) 「か」(並立助詞と副助詞) など

   - 最も多い誤り。
   - 学習量をさらに増加することによって、若干の改善は見込める。

   (b) 異品詞間の誤り
   「で」(判定詞と格助詞) 「ない」(形容詞と助動詞) など

   - 言語モデルの改良 (bi-gram から tri-gram や 可変 n-gram へ) による解決法も可能性あり。

参考として、以下に誤った文の例を示す。

分割誤り

(正解:「に (格助詞)」+「なる (本動詞)」、出力:「になる (補助動詞)」)[1]
おそらく午後三時ごろの到着 になる と思います。
はいいろいろとお世話 になり ますが一つお願いいたします。
そして京福嵐山線に乗り換え になっ てください。
じゃあ一人五十二ドルっていう計算 になり ますね。

(正解:「週 (普通名詞)」+「間 (普通名詞)」、出力:「週間 (普通名詞)」)[2]
で日程はですね八月六日の土曜日から七日間一 週間 ほど沖縄に滞在します。
わたしと妻の二人です日程は八月六日土曜日から一 週間。
これから一 週間 ぐらいはまだ日本にいるんですが。

[1]現在の形態素仕様では、「サ変名詞に後接する『になる』を補助動詞とする」とある。
[2]現在の形態素仕様では、数詞が前接する場合は「週」+「間」、しない場合は「週間」となっている。

品詞付与 (助詞細分類間) 誤り

(正解: 格助詞、出力: 並立助詞)
で環状線外回り と 内回り と ございますので...
カード番号 と 有効期限はどのようになっていますでしょうか。
あす主人 と 二人でセントラルパークを三時間ほどかけて散策したいんです。
近鉄の京都駅はＪＲの京都駅 と 同じ建物になりますが。

(正解: 並立助詞、出力: 副助詞)
何 か 特に問題はございませんか。
他に何 か 薬を併用なさっていらっしゃいますか。
はい昼の公演でしたら何枚 か チケットございます。
友人からなん か ニューヨークで本のお祭があるって聞いたんですけれど。
一人当り一万円 か ちょっとそれを越えるぐらいなんですけども。

品詞付与 (異品詞間) 誤り

(正解: 判定詞、出力: 格助詞)
そうです駅名はアメリカ自然史博物館 で そこで下車します。
カバーチャージは十ドル で 最低飲み物を一杯ご注文いただきます...
... 一番小さいのが七十五セント次に大きいのが一ドル で 一番大きいのが...
はいとても有名 で とても人気のあるレストランです。
はいこれは非常においしいサラダ で お野菜にシーフードそれに...
朝五時から十一時までは朝食メニュー で 和食と洋食の両方がございます。

(正解: 形容詞、出力: 助動詞)
... すみませんが太秦か嵐山辺りでどこかいい所 ない でしょうか。
ニューヨークだったらどこかに ない とおかしいですよねえ。
そうですかでも氷が ない んですよねえ。

3.2.2 TDMT にとっての重要度

前節に示した誤りは、仕様に照らして(コーパスとの一致による評価)の誤りであったが、実際このうちの半分は翻訳システム(TDMT)ではその相違を折り込み済みであるか、またはTDMTにとって必要以上の品詞細分化であるため、翻訳の結果には影響しない誤りである。そもそも本形態素解析システムは、TDMTで必要な処理を行なうための前処理として解析を行なっているので、形態素解析の誤りはすべて同等に扱う必要はなく、誤りのうちでもTDMTで悪影響を及ぼす誤りをより重視すればよい。

TDMTにとって特に影響の大きいと考えられている形態素誤りは、分割誤りと、内容語と機能語の判定誤りの二点である。前節の誤り分析によると、幸いなことに前者の誤りは少なく、後者も「ある」「ない」などの一部の特殊な語を除けばほとんど見られない。このことから、本形態素解析プログラムはTDMTが要求する形態素解析の最低水準を満たしていると考えることができる。しかしまだ、言語モデルの改良などにより解析精度向上の余地がある。

3.3 (参考) 従来の書き言葉解析プログラムとの性能比較

JUMAN と ALT-JAWS は書き言葉を対象とした、公開され広く利用されている高精度の形態素解析プログラムである。書き言葉用に開発されたこれらのプログラムで話し言葉を処理することには無理があるし、逆も同様である。また、テストコーパスや品詞体系、辞書サイズの相違もあり、既発表の性能 (数値) を直接比較することにも無理がある。しかし、ここではあえて参考として上記両システムと ATR のシステムについてデータを示す。

JUMAN

研究用として現在最も広く使用されている形態素解析プログラムである、JUMAN (京大、奈良先端大) の定性的な特徴を以下に示す。

- 全体的に、ひらがな (の連続) に対して弱い。
- (特に複合語、接辞に対して) 解析のゆれが見られる。
- 形容動詞は「名詞＋だ」と、サ変動詞は「名詞＋する」と解析する。
- 3.1節に示したような表現は苦手である。

JUMAN の解析精度は、文献 [Tok96] によると形態素分割精度[3]99% 及び品詞付与精度[4] 93% とあり、[山地 96] では EDR コーパスの 25000 文に対して 99.68% の単語分割精度であったという報告がある。

大雑把な比較として解析精度に関してほぼ JUMAN と ATR のシステムは同等であると考えられる。

ALT-JAWS

NTT が開発している機械翻訳システム ALT-J/E で使用している形態素解析システム ALT-JAWS については、965 文の新聞リード文に対して実験を行なった結果、形態素分割のみの適合率が 99.86%、品詞付与も含んだ正解率が 99.5% 程度と報告されている ([白井 95])。

一般に形態素解析で誤りやすいひらがな列は、新聞記事よりも話し言葉 (の書き起し) に多く出現することを考慮に入れると、ALT-JAWS の精度も本システムの精度と大きな差はなく、同等の解析精度であると考えることができる。

[3]入力文を形態素に分割した時の、区切り位置のみを考慮した場合の正確さ。

[4]形態素に分割した位置の正確さと付与した品詞の正確さの両者を考慮した精度。

本システムとの比較

以上の各文献の記述をまとめたものを表 3.1に示す。

表 3.1: 従来の書き言葉用システムの解析精度: まとめ

	形態素分割	品詞付与
JUMAN[Tok96]	99%	93%
JUMAN[山地 96]	99.68%	—[5]
ALT-JAWS[白井 95]	99.86%	99.5%

一方比較対象として、今回の実験 (32561 文の 10 分割 cross validation) での本形態素解析プログラムの分割精度及び品詞付与精度を以下に示す。

表 3.2: 本プログラムの解析精度 (32561 文)

Top-n	形態素分割 (Pcn.)	品詞付与 (Pcn.)
1	99.52%	98.88%
5	99.95%	99.84%
10	99.97%	99.88%

[5]文献に記述なし。

第 4 章

まとめ

日英韓各言語の形態素解析プログラムの評価結果について述べた。本報告書の内容をまとめると以下のようになる。

- 品詞と単語の混合 n-gram と ATR の対話コーパスを採用したことにより、話し言葉にも対応したトップクラスの解析精度 (日本語の場合、単語分割精度 99.5%、品詞付与精度 98.9%) であることが検証できた。
- 速度面でも、SPARCstation 10 で一文あたり日本語で 0.2 秒程度と、ほぼ満足できる。
- ほぼ同一の機構で日、英、韓の三ヶ国語を取り扱うことができた。
- 言語モデルやアルゴリズムの改良による性能向上の余地はある。
- 各言語について：
  - 英語解析は精度、処理時間共に全く問題なし。
  - 日本語解析は、現在最高水準にあると考えられている JUMAN や ALT-JAWS の二つのプログラムと比較してほぼ同等の精度であった。
  - 韓国語解析はまだ精度向上の余地が大きい。

また、形態素に関して今後取り組むべき主な課題は以下の通り。

1. 未知語処理
2. 韓国語の精度向上
3. 品詞体系の再検討

参考文献

[Nag94] NAGATA, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm, In *Proc. of Coling 94*, pp. 201–207 (1994).

[Tok96] TOKUNAGA, T. and IWAYAMA, M.: Word-based vs. Character-based indexing: An experimental study on Japanese text representation for text categorization, In *Proc. of the workshop on Information Retrieval with Oriental Languages*, pp. 73–78, KORDIC (1996).

[丸山 94] 丸山宏, 荻野紫穂：正規文法に基づく日本語形態素解析, 情報処理学会論文誌, Vol. 35, No. 7, pp. 1293–1299 (1994).

[山地 96] 山地治, 黒橋禎夫, 長尾眞：連語登録による形態素解析システム JUMAN の精度向上, 年次大会発表論文集, 第 2 回, pp. 73–76, 言語処理学会 (1996).

[中川 88] 中川聖一：確率モデルによる音声認識, 電子情報通信学会 (1988).

[田代 96] 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).

[白井 95] 白井諭, 横尾昭男, 池原悟, 奥山信輔, 宮崎正弘：多段解析法による日本語形態素解析の精度, 全国大会講演論文集, 第 50 回, IR-2, pp. 3/37–38, 情報処理学会 (1995).