形態素解析辞書のベンチマークテスト
—IPAdic・NAIST-jdic・UniDic のジャンル別精度比較—

小木曽智信[†]，小椋秀樹[†]，小磯花絵[†]，宮内佐夜香[†]，渡部涼子[†]，伝康晴[‡]

[†]国立国語研究所，[‡]千葉大学文学部

1. はじめに

発表者らは形態素解析辞書 UniDic の開発を行っている。UniDic は、

1. 「短単位」という揺れが少ない斉一な単位を見出し語に採用している
2. 語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを与えることができる
3. 話し言葉のテキストの解析に対応しているほか、アクセントや音変化の情報を付与することができ、音声処理の研究に利用できる

といった特徴を持つ解析辞書であり、国立国語研究所を中心に構築が進む『現代日本語書き言葉均衡コーパス』（BCCWJ）の構築にも利用されている（伝ほか 2007）。

多様なジャンルのテキストからなる BCCWJ の解析に対応するため、UniDic は多ジャンルのテキストから見出し語の追加を行っており、その数は語彙素（国語辞典の見出し相当）で 15.7 万語、書字形（表記形）で 23 万語に達している（UniDic-1.3.12）。UniDic では、学習用コーパスにも多ジャンルのテキストを利用しているため、専ら新聞記事のデータを利用してきた従来の形態素解析辞書に比べ、多様なテキストを高い精度で解析することが可能になっている。

本発表は、UniDic が多ジャンルのテキストを高い精度で解析できることを確認するため、その解析精度を代表的な形態素解析辞書である IPAdic、その後継にあたる NAIST-jdic と比較する試みである。3 つの解析辞書により、新聞・文学作品・ブログの 3 つのジャンルのデータを解析して、(1)読み推定，(2)品詞推定，(3)読みと品詞の推定の 3 つのタスクについて、解析精度を比較する。

2. 調査の概要

2.1. 対象とする形態素解析辞書

調査の対象とした形態素解析辞書は**表 1** のとおりである。3 種とも、ChaSen 版と MeCab 版が利用可能だが、今回は全て MeCab 版を用いた。解析器は mecab-0.98 を使用した。解析器の設定は、各辞書に付属の設定ファイルをそのまま利用した。

表 1　調査対象の形態素解析辞書

解析辞書	バージョン
IPAdic	mecab-ipadic 2.7.0-20070801 （mecab-win32　0.98 付属）
NAIST-jdic	mecab-naist-jdic-0.6.1-20090630
UniDic	UniDic-mecab 1.3.12

2.2. 対象とするテキスト

解析対象のテキストとして、現代日本語として広く読まれている、新聞・文学作品・ブログの 3 ジャンルを対象とした。

新聞のデータは『CD-毎日新聞 2007 データ集』を用いた。文学作品のデータは『CD-ROM 新潮文庫の 100 冊』をテキスト化したものを用いた。ブログについては、BCCWJ の構築に用いられる「Yahoo!ブログ」のデータの一部を用いた。Yahoo! ブログのデータは、現時点では UniDic の見出し語追加に一切利用していない。したがって、いずれも、3 種の形態素解析辞書にとって未知のデータである。

調査対象のサンプルは、これら 3 ジャンルのデータから不要な記号類やタグを除去し、改行と句点を基準にして文単位に区切ったものから、各ジャンルにつき約 3.5 万字分（1000 文。ただし文が短いブログでは 1200 文）を文単位でランダムサンプリングしたものである（**表 2**）。

なお、特にブログでは文の途中に改行が入り、文の断片がサンプルに紛れ込むことがある。このような場合は手作業で調査対象から除いた。

表 2　調査対象のテキストデータ

ジャンル	データ	サンプル文数	サンプル文字数
新聞	毎日新聞 2007 年度版 （約 118.1 万文）	1000	36364
文学作品	新潮文庫の 100 冊 （約 59.2 万文）	1000	35453
ブログ	Yahoo!ブログ （約 19.3 万文）	1200	34209

2.3. 評価するタスク

辞書ごとに品詞体系や見出し語の長さが異なるため、解析結果を直接比較して評価することは難しい。そこで、(1) 読み推定，(2) 品詞推定，(3) 読みと品詞の推定の 3 つのタスクを想定し、それぞれのタスクにおける精度を比較することとした。

(1) 読み推定は、形態素解析を用いてテキストにふりがなを付与することを想定したタスクである。ここでは、語の境界が正しいかどうかは問わず、読みとして正しいカタカナが出力されていれば正解とみなした。未知語として出力された語であっても、ひらがな・カタカナ表記の語の場合は正解とした。

(2) 品詞推定は、形態素解析を用いて特定の品詞の語を抜き出すことを想定したタスクである。解析辞書により品詞体系が異なるため、辞書間の差異が少ない最上位の品詞区分（品詞大分類）までを評価対象とした。なお、全ての辞書について、解析器による未知語の品詞推定は行わず、未知語は品詞「未知語」として出力し、品詞誤りとして取り扱った。

(3) 読みと品詞は、上記(1)と(2)の組み合わせであり、双方が正しい場合を正解とした。

2.4. 正誤判定の基準

正誤の判定にあたっては、特定の辞書に不利になることがないように十分配慮しつつ、次のような基準を設けて判定を行った。判定にあたっては 2 名の作業者によるクロスチェックを行った。

- 名詞連続などにおける語の長さの揺れは誤りとしない。
  消費/税込み，銘/板，消しゴム/印
- 解釈の複数考えられる文法事項等の解析のゆれは誤りとしない。
  届いたのよ　「の」終助詞・準体助詞
  お勧めです　「勧め」名詞・動詞連用形
  誰とでも　「で」格助詞・助動詞「だ」連用形
- 品詞評価は品詞大分類（14～15 カテゴリ）までとするが、固有名詞については姓・名・地名の別まで区別する。
- 品詞大分類が辞書によってゆれる名詞的接尾辞は、名詞・接尾辞のいずれであっても正解とする。
- アルファベット表記の語については、日本語の文章中で一般に使用されるものは正しく品詞づけが行われていなければ誤りとし、一般的でないものや外国語単語は評価対象外とする。
  ○誤り：ＰＣ・ＦＷ・ＳＡＴ・ＫＡＧＯＭＥ・ＭＢＳ　等
  ○対象外：Ｂｅａｕｔｉｆｕｌ　ｌｉｆｅ・ＺＥＰＰ　等
- ブログにおける誤字や顔文字、いわゆるギャル文字、URL 等は調査対象外として集計から外す（顔文字は別途評価）。

UniDic が詳細な単位認定規程（小椋ほか 2009）をもつのに対して、他の辞書では必ずしも厳密な基準が明示されていない。正誤の判断がつかない場合には誤りとしなかったため、結果としてやや UniDic に厳しい判定となっている可能性がある。

3. 評価結果

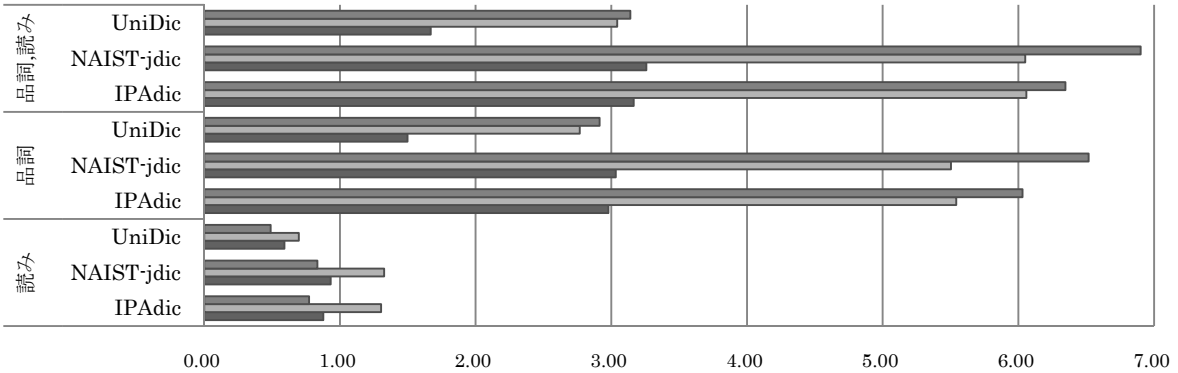
3.1. 解析誤りの集計結果

各辞書・ジャンル・タスクごとに解析誤りの数を調査した結果を**表 3** に示す。評価は出力された語を単位として行っており、表中の文字数は誤りを含む語の文字数を集計したものである。

品詞誤りには未知語として出力された語数も含んでいる。UniDic の未知語の数は、他の辞書の 1/5～1/4 となっている。

表 3　解析誤りの集計結果

		IPAdic		NAIST-jdic		UniDic	
		語数	文字数	語数	文字数	語数	文字数
新聞	出力語数	23358	36364	23350	36364	23826	36364
	読み誤り	214	319	227	339	155	215
	品詞誤り	447	1083	449	1103	289	545
	（うち未知語）	185	771	150	728	38	219
	いずれかの誤り	495	1151	506	1185	333	607
文学作品	出力語数	22620	35453	22606	35453	23106	35453
	読み誤り	379	462	383	470	205	247
	品詞誤り	945	1965	938	1951	526	981
	（うち未知語）	223	833	201	764	40	165
	いずれかの誤り	1102	2148	1104	2145	606	1079
ブログ	出力語数	21578	33910	21471	33910	21930	33910
	読み誤り	226	262	238	283	124	166
	品詞誤り	870	2045	922	2210	535	988
	（うち未知語）	529	2101	628	2305	146	633
	いずれかの誤り	967	2152	1038	2340	593	1065



	読み			品詞			品詞,読み		
	IPAdic	NAIST-jdic	UniDic	IPAdic	NAIST-jdic	UniDic	IPAdic	NAIST-jdic	UniDic
ブログ	0.7726	0.8346	0.4895	6.0307	6.5173	2.9136	6.3462	6.9006	3.1407
文学作品	1.3031	1.3257	0.6967	5.5425	5.5031	2.7670	6.0587	6.0503	3.0435
新聞	0.8772	0.9322	0.5912	2.9782	3.0332	1.4987	3.1652	3.2587	1.6692

図 1　文字数ベースの誤り率（%）

3.2.　文字ベースの誤り率

表 3 をもとに、解析辞書の見出し語の長さに左右されずに比較が可能な文字数をベースとして、各解析辞書の誤り率を集計した結果を**図 1** に示す。

全てのジャンル、全てのタスクにおいて、UniDic は他の辞書のほぼ半分程度の誤りであり、圧倒的に精度が高い。IPAdic と NAIST-jdic はおおよそ同じ誤り率だが、全体として IPAdic の方が誤りが少ない。ジャンル別に見ると、どの辞書においても新聞が他のジャンルの半分程度の誤り率であり、解析しやすいテキストであることがわかる。新聞では表記や語法が統制されているのに対し、文学作品やブログでは多様な表現が用いられるためだと考えられる。ブログが、読み推定でのみ誤りが少ないのは仮名書きの語が多いためである。なお、語をベースとして誤り率を比較した場合にも図 1 と同様の傾向がみられる。

3.3. 各解析辞書の精度

各サンプルについて、各解析辞書の品詞体系において正しいと考えられる単位分割を行ったときの語数を調査し、この結果と表 3 をもとに、各辞書の解析精度を計算した結果を**表 4・図 2** に示す。各辞書において正しいと考えられる語数（Precision の分母）は、各辞書の出力結果をもとに、切り直し箇所が最も少なくなるように判定したものである。そのため、品詞体系がほぼ同じ IPAdic と NAIST-jdic とでも語数は異なる。

ジャンル別に品詞認定のレベルを見ると、新聞ではいずれの辞書も 98%近い精度を示すものの、他のジャンルでは、IPAdic と NAIST-jdic は 95～96%まで下がる。一方、UniDic は 98%近くを維持しており、多様なジャンルのテキストに対応できていることがわかる。

3.4. 顔文字の解析

今回の調査では顔文字は対象外としたが、UniDic には顔文字が見出し語として登録されており、「補助記号-ＡＡ-顔文字」という品詞が付与される。今回のサンプルには 55 カ所、299 文字分の顔文字が確認されたが、そのうち 22 箇所で正しく顔文字として解析されていた。残る 33 カ所については、顔文字を見出し語に持たない他の辞書と同様、ほとんどが単字の補助記号（一部は未知語ないし誤解析）として解析されている。

4.　おわりに

調査した全てのジャンル、全てのタスクにおいて UniDic の解析精度が他の辞書にまさっていることが確認された。特に文学作品とブログでは大きな差をつけている。UniDic は、語彙素・語種・アクセント型など他の辞書よりも多くの情報を付与することが可能であり、応用範囲は広い。今後、さらに多くの研究で利用されることを期待したい。

参考文献

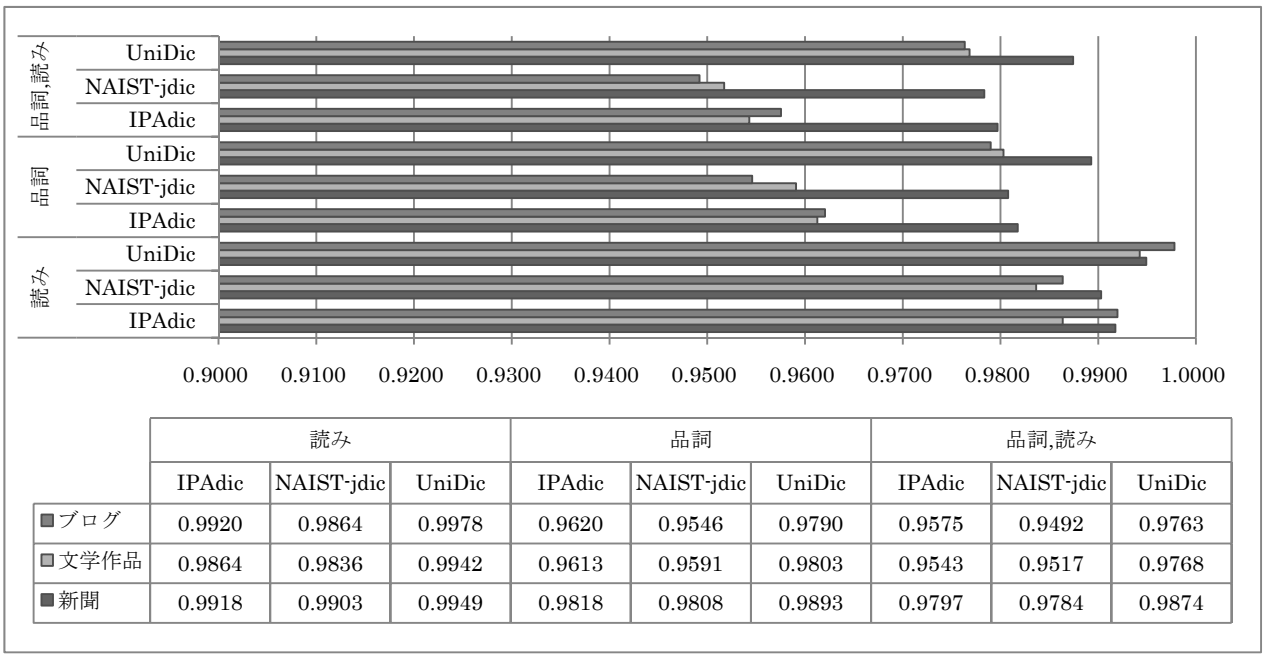
【論文・報告書】
・伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のた

めの言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22　pp.101-123
・小椋秀樹・小磯花絵・冨士池優美・原裕（2009）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集改定版』(LR-CCG-08-03)
【テキストデータ】
・『CD-毎日新聞 2007 データ集』毎日新聞社／日外アソシエーツ
・『CD-ROM 新潮文庫の 100 冊』新潮社
・『Yahoo!ブログ』http://blogs.yahoo.co.jp/
【解析器】
・MeCab　http://mecab.sourceforge.net/
【解析辞書】
・形態素解析辞書 UniDic　http://download.unidic.org
・IPAdic legacy　http://sourceforge.jp/projects/ipadic/
・NAIST Japanese Dictionary
http://sourceforge.jp/projects/naist-jdic/

付記
本発表は科研費・特定領域研究「日本語コーパス」による成果の一部を含むものである。

表 4　各解析辞書の精度

		IPAdic			NAIST-jdic			UniDic		
		新聞	文学作品	ブログ	新聞	文学作品	ブログ	新聞	文学作品	ブログ
出力語数		23358	22620	21578	23350	22606	21471	23826	23106	21930
語数（正解）		23315	22477	21472	23349	22579	21582	23758	22961	21778
読み	正解数	23144	22241	21352	23123	22223	21233	23671	22901	21806
	Precision	0.9927	0.9895	0.9944	0.9903	0.9842	0.9838	0.9963	0.9974	1.0013
	Recall	0.9908	0.9832	0.9895	0.9903	0.9831	0.9889	0.9935	0.9911	0.9943
	F 値	0.9918	0.9864	0.9920	0.9903	0.9836	0.9864	0.9949	0.9942	0.9978
品詞	正解数	22911	21675	20708	22901	21668	20549	23537	22580	21395
	Precision	0.9809	0.9582	0.9597	0.9808	0.9585	0.9571	0.9879	0.9772	0.9756
	Recall	0.9827	0.9643	0.9644	0.9808	0.9597	0.9521	0.9907	0.9834	0.9824
	F 値	0.9818	0.9613	0.9620	0.9808	0.9591	0.9546	0.9893	0.9803	0.9790
品詞,読み	正解数	22863	21518	20611	22844	21502	20433	23493	22500	21337
	Precision	0.9788	0.9513	0.9552	0.9783	0.9512	0.9517	0.9860	0.9738	0.9730
	Recall	0.9806	0.9573	0.9599	0.9784	0.9523	0.9468	0.9888	0.9799	0.9798
	F 値	0.9797	0.9543	0.9575	0.9784	0.9517	0.9492	0.9874	0.9768	0.9763



	読み			品詞			品詞,読み		
	IPAdic	NAIST-jdic	UniDic	IPAdic	NAIST-jdic	UniDic	IPAdic	NAIST-jdic	UniDic
■ブログ	0.9920	0.9864	0.9978	0.9620	0.9546	0.9790	0.9575	0.9492	0.9763
■文学作品	0.9864	0.9836	0.9942	0.9613	0.9591	0.9803	0.9543	0.9517	0.9768
■新聞	0.9918	0.9903	0.9949	0.9818	0.9808	0.9893	0.9797	0.9784	0.9874

図 2　解析辞書のジャンル別精度（F 値）