言語資源・ツールの研究利用状況に関する調査報告

齋藤 邦子*1*2　井佐原 均*1*3　下畑 さより*1*4　白井 清昭*1*5　光石 豊*1*6

*1 (社)電子情報技術産業協会 知識情報処理技術専門委員会 言語資源分科会

*2 NTT サイバースペース研究所, *3 豊橋技術科学大学, *4 沖電気工業株式会社,

*5 北陸先端科学技術大学院大学, *6 株式会社富士通研究所

nlp_portal@nlp.kuee.kyoto-u.ac.jp

1　はじめに

近年、自然言語処理の研究において、新聞記事、注釈付コーパス、辞書、シソーラスなど、様々な言語資源が活用されている。JEITA言語資源分科会では自然言語処理研究の発展に資する言語資源の発見と普及を目指した様々な活動を行ってきた。例えば、一般に入手可能な言語資源・ツールのカタログ情報を言語情報処理ポータル[1]で公開している。

言語資源の普及を促進したり、新しい言語資源を開発するためには、どのような言語資源がよく使われているのか、研究に必要とされている言語資源は何かなど、言語資源の利用状況を把握することが重要である。このような背景から、我々は2006年から2010年の言語処理学会年次大会の発表論文を対象とし、これらの論文における言語資源・ツールの利用状態を調査した。本稿ではその調査結果を報告する。

2　調査の概要

まず、調査方法の概要について述べる。2006年から2010年の言語処理学会年次大会の発表論文から言語資源・ツールの名称を文字列マッチングにより検索する。言語資源名が見つかったとき、その言語資源がその論文で研究に使用されているとみなし、言語資源の使用回数を集計する。さらに、発表者が申込時に入力した「該当分野」の情報から論文の研究分野を特定することで、それぞれの研究分野における言語資源の使用回数も集計する。これらの集計結果を主に以下の2つの観点から分析する。

(1) よく使用される言語資源は何か?　また年によって使用状況に変化があるか?

(2) 研究分野によって使用される言語資源に違いがあるか? 特定の研究分野でよく使われる言語資源があるか?

[1] http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/

表1．調査対象一覧

(a) 言語資源

ジャンル	項目数	項目名(一部抜粋)
新聞	6	朝日新聞, 日経新聞, 毎日新聞, 読売新聞
平文	12	青空文庫, Basic Travel Expressions, British National Corpus
注釈	11	NTCIR テストコレクション, 京都テキストコーパス, 現代日本語書き言葉均衡コーパス(BCCWJ), 日本語話し言葉コーパス(CSJ)
シソーラス	6	分類語彙表, 日本語 WordNet, 日本語語彙大系
辞書	26	EDICT, EDR 辞書, IPAL 辞書,UniDic, 英辞郎
音声	16	ATR 多数話者音声 DB, 日本音響学会新聞記事読み上げコーパス
その他	12	Web 日本語 N グラム,Wikipedia

(b) ツール

ジャンル	項目数	項目名(一部抜粋)
形態素解析	5	JUMAN, 茶筌, 和布蕪
パーザ	5	KNP, MSLR パーザ, SAX, 南瓜
その他	21	Julius, Moses, SRILM, SVMlight, Tagrin, TinySVM, TSUBAKI, YamCha

2.1　調査対象となる言語資源・ツール

今回の調査では、言語情報処理ポータルで紹介している言語資源のカタログ項目をベースとし、更にカタログには掲載されてはいないが重要と思われるものを追加して調査対象とした。その際、全体をいくつかのジャンルにグループ分けをしており、言語資源は{新聞、注釈、平文、辞書、シソーラス、音声、その他}の7つのジャンルで総計89項目、ツールは{形態素解析、パーザ、その他}の3つのジャンルで総計31項目を調査対象とした。

ジャンルの一覧及び今回の調査で使用回数の多かった各ジャンルの代表的な項目名を表 1 に挙げる。

2.2　抽出ツール

論文から言語資源を抽出するツールを作成し

た。まず、論文のPDFファイルから、pdftotext[2]を用いてテキストを抽出する。次に、言語資源の項目名を全文検索し、1 回以上ヒットした場合、その言語資源は論文で利用されているとみなして抽出する。ただし、カタログに記載されている言語資源の正式名称だけでは論文の表記とマッチしないケースが多い。そこで、あらかじめ人手で正式名から想定される表記揺れを追加登録しておき、これらも検索キーとして利用することで検索漏れを軽減させた。追加した表記揺れの数は73 語であり、元々の正式名称の総数（89+31）と併せて 193 語で自動抽出を行った。

2010 年の年次大会の論文からツールによって抽出された言語資源の一部（106 件）を人手でチェックしたところ、抽出誤り（実際には言語資源を使用していないのに抽出されたケース）が 9 件あり、抽出精度はおよそ 92%であった。このことから今回の集計結果は利用実態を概ね正しく反映できていると考えている。ただし分析の過程で、ある研究分野に誤抽出が多く確認されたが、これについては 3.2.1 節で詳しく述べる。

2.3 研究分野

本調査では、発表者が発表申込みをする際に申請する「第一該当分野」の情報を利用して、論文の研究分野を特定する。ただし、言語処理学会が設定する分野分類は本調査には詳細な体系となりすぎるため、この分類体系を独自に 9 つの研究分野に分類しなおした。なお、言語処理学会の分野分類は開催年によって若干変更されることがあるため、それぞれの年度の体系に応じて同等の分野変換を行った。表 2 に一例として 2010 年開催の言語処理学会の分類と、本調査で利用した研究分野分類の対応を示す。

表 2. 分析に利用する研究分野体系

本調査の分野分類		NLP2010 申込時の分野分類
lin	言語学	A. 言語学・言語分析　全て
ana	言語解析	B(2)形態素解析, B(3)構文解析 B(4)意味解析, B(8)文脈処理
cor	言語資源・ 語彙・辞書	B(1)語彙・辞書 B(9)言語資源・コーパス
gen	要約・生成・ 言い換え	B(7)生成, B(10)言い換え C(4)要約
cla	分類	B(12)文書分類
dia	対話・音声	C(3)対話, C(9)音声言語処理
ext	マイニング・ 抽出	B(5)固有表現解析, B(6)評判・感情解析 B(11)知識獲得, C(5)情報抽出 C(8)テキスト・データマイニング
mt	機械翻訳	C(1)機械翻訳
sys	検索・QA・応 用システム	C(2)情報検索, C(6)質問応答 C(7)Web 応用, C(10)教育応用

[2] http://www.foolabs.com/xpdf

3 調査結果と分析

3.1 言語資源・ツール別の使用傾向

まず、調査対象とした言語資源・ツールのうち、使用回数の経年変化を調査した。表 3 に結果を示す。各ジャンルで利用回数合計が最大のものを青字で、近年使用回数が増加しているものを赤字で示した。なお、過去 5 年間の利用回数合計が 3 回以下の項目は省略したが、右端の「ジャンル別総計」欄は 5 年間の全項目での使用頻度をジャンル別に合算したものである。

ジャンル別に 5 年間の使用頻度総計をみると、言語資源では、新聞 24%、平文 7%、注釈 27%、シソーラス 14%、辞書 16%、音声 1%、その他 11%であり、注釈、新聞、辞書の利用頻度が高い。またツールでは、形態素解析 48%、パーザ 25%、その他 27%と、形態素解析ツールの利用頻度が極めて高い。

更にジャンルごとに詳細を見ると、際立って高頻度で利用されたり、近年急速に利用が伸びている項目がある場合があったので幾つか紹介する。

【新聞】圧倒的に使用回数が多いのは毎日新聞で新聞ジャンルの 66%を占めるが、日経新聞・読売新聞もコンスタントに利用されている。毎日新聞の利用が高いのは、特に 94・95 年のデータに様々な情報が付与された注釈付コーパスが存在しているためと思われる。

【注釈】NTCIR テストコレクションが注釈ジャンルで 33%と最多であるが、京都コーパス・CSJ も堅調である。また、BCCWJ の利用がここ 2 年で急速に伸びている。

【シソーラス】語彙大系が 57%と最も活用されており、分類語彙表が 32%と続いている。日本語 WordNet が今後伸びそうな兆しが見受けられる。

【辞書】EDR 辞書、英辞郎の利用が高い中、UniDic の伸び率が高い。

【その他（言語資源）】Wikipedia の利用が著しく伸びている。Wikipedia については次節の分析でも注目することとする。

【形態素解析】茶筌が 47%と半数を占めるが、近年和布蕪の利用も伸びている。

【その他（ツール）】TinySVM が最多であるが、Moses、TSUBAKI、SRILM の伸びも近年は高い。

表 3. 言語資源・ツールの使用頻度

言語資源		2006	2007	2008	2009	2010	total	ジャンル別総計
新聞	朝日新聞	0	1	1	1	0	3	
	日経新聞	5	10	4	10	8	37	
	毎日新聞	35	31	19	23	22	**130**	
	読売新聞	6	9	3	4	5	27	197
平文	**Basic Travel Expressions Corpus**	7	4	2	3	4	**20**	
	British National Corpus	3	6	3	2	3	17	
	青空文庫	5	4	3	4	0	16	
	特許公報類 CD-ROM	1	1	1	2	3	8	61
注釈	EDR 日本語コーパス	2	0	2	2	0	6	
	IREX 公開データ・ツール	2	1	7	4	3	17	
	NTCIR テストコレクション	15	20	12	14	14	**75**	
	京都テキストコーパス	11	10	10	8	9	48	
	現代日本語書き言葉均衡コーパス(BCCWJ)	0	4	5	**12**	**18**	39	
	日本語話し言葉コーパス(CSJ)	9	11	10	5	6	41	229
シソーラス	分類語彙表	8	11	6	8	6	39	
	日本語 WordNet	0	0	0	1	**9**	10	
	日本語語彙大系	25	14	8	10	8	**65**	114
辞書	EDICT	1	4	3	2	2	12	
	EDR 日本語単語辞書	1	2	0	1	0	4	
	EDR 概念辞書	1	1	0	1	2	5	
	EDR 辞書	9	13	6	4	3	**35**	
	IPAL 辞書	1	2	2	2	1	8	
	MUST1: 日本語複合辞用例データベース	2	2	1	4	3	12	
	UniDic	1	1	**4**	**8**	**4**	18	
	英辞郎	6	6	3	7	4	26	
	ライフサイエンス辞書	0	0	0	2	2	4	131
音声	日本音響学会 新聞記事読上音声コーパス	0	1	2	2	0	5	6
その他	Web 日本語 N グラム第 1 版	0	0	1	1	2	4	
	Wikipedia	4	9	**24**	**27**	**26**	**90**	96
ツール		2006	2007	2008	2009	2010	total	
形態素解析	JUMAN	16	26	26	19	19	106	
	茶筌	44	45	27	30	32	**178**	
	和布蕪(MeCab)	6	15	23	14	**34**	92	376
パーザ	KNP	15	17	20	13	10	75	
	南瓜(CaboCha)	23	29	20	19	31	**122**	200
その他	Julius	1	2	0	3	3	9	
	Moses	0	0	**6**	**13**	**16**	35	
	SRILM	0	1	**4**	**8**	**12**	25	
	SVMlight	1	3	1	3	4	12	
	Tagrin	1	1	2	0	1	5	
	TinySVM	10	10	14	12	8	**54**	
	YamCha	9	9	3	2	1	24	
	検索エンジン基盤 TSUBAKI	0	3	**6**	**9**	**11**	29	
	節境界解析ツール CBAP	5	2	3	2	2	14	209

3.2 研究分野別の使用傾向

続いて研究分野別の使用傾向について詳細を調査した。この調査では、ツールは対象外としている。更に、新聞、平文、注釈、シソーラス、辞書、音声の 6 ジャンルと、その他ジャンルの Wikipedia に限定した時の利用状況を分析した。

3.2.1 論文あたりの使用回数比較

まず、各研究分野について、注目する言語資源の使用回数が 1 つの論文あたりどのくらいになるかを調査した。

図 1(a)に結果を示す。lin(言語学)では全体に使用回数が少ないものの、平文に関しては他の分野と遜色なく使用されている。また新聞も平文と同程度に利用されている。一方、ana(言語解析)、cor(言語資源・語彙・辞書)、gen(要約・生成・言い換え)では使用回数が高い。cla(文書分類)は使用回数がやや低いが、新聞・平文・Wikipedia をあわせたプレーンテキスト相当のデータは他の分野より高い頻度で利用されている。dia(対話・音声)は分野として該当しているはずの音声 DB も思いのほか利用が少ない。またこのケース

に限らず、他の分野でも使用回数が少ない事例については、現在のカタログにある資源では目的に合わず、自分たちでデータ構築をしているケースも多いと予想され、今後の言語資源に対するニーズの発掘に役立てたい。

この調査結果全体を眺めると、言語資源のジャンルによって利用度の高い研究分野が異なることがわかる。近年急速に利用が伸びているWikipediaはcla・sys(検索・QA・応用システム)、新聞はgen、平文はmt(機械翻訳)、シソーラスはana・ext(マイニング・抽出)、辞書はmt(機械翻訳)、注釈はanaでよく利用される。なお、corは新しい言語資源の開発に関する論文が多いが、それでも注釈の利用頻度が高い。これは我々の当初の予想と異なっていたため、corの結果を人手で全てチェックしたところ、幾つかの誤抽出が確認された(注釈は 53 件中 14 件、シソーラスは 24 件中 3 件、辞書は 33 件中 7 件)。これは、corの研究目的が新たな言語資源の構築である場合、既存の注釈・辞書データが関連研究として記載されるためである[3]。

3.2.2 使用回数の割合比較

続いて、各研究分野について、言語資源の使用回数割合をジャンル別に比較した(図 1(b))。以下、表には現れないが、各分野で利用の高かった言語資源についてもあわせて紹介する。

【lin】注釈の割合が高く、特に CSJ・BCCWJ の使用が多かった。また音声も利用される。

【ana】同じ注釈でも京都コーパス・BCCWJ の使用が多く、またシソーラスや辞書として語彙大系、EDR 辞書・UniDic も頻度が高かった。

【cor】前述の通り一部誤抽出が確認されたが、全体にどの資源も満遍なく使われる傾向にある。

【gen】新聞の割合が約 4 割で、特に毎日新聞が多い。また注釈では NTCIR が良く利用されている。

【cla】辞書の使用が殆どなく、毎日新聞に限らず新聞全般、また Wikipedia の利用が目立つ。

【dia】音声の他、注釈での CSJ や毎日新聞、Wikipedia の利用が高い。

【ext】新聞では毎日の他に日経新聞も良く使われる。また注釈では IREX・NTCIR の利用が高い。

【mt】辞書の利用が高いが、英辞郎や EDR 辞書を始めとして様々な辞書が利用されていた。また新聞も毎日新聞・読売新聞ともによく利用される。

【sys】毎日新聞と Wikipedia の利用が目立ち、また注釈として NTCIR が良く利用されている。辞書も mt と同様に EDR 辞書・IPAL 辞書・UniDic や英辞郎など様々な辞書が幅広く使われていた。

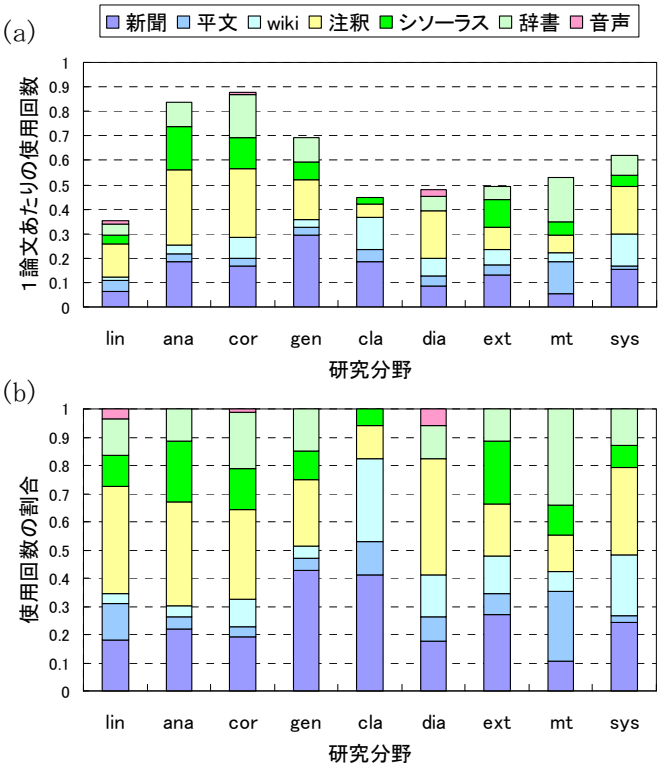

図 1. 研究分野別の言語資源使用状況
(a)1 論文あたりの使用回数 (b)使用回数の割合

4 まとめ

過去 5 年間の言語処理学会年次大会発表論文集を元に、言語資源・ツールの利用状況を調査した。一方、我々はカタログ未掲載の言語資源や、研究者が独自に作成した評価データの調査にも着手している。今回の調査結果と併せて、研究者にとってニーズの高い言語資源は何かを明らかにし、それらを新たに発掘・構築することで、言語資源利用のさらなる促進に貢献したい。

謝辞 過去の年次大会の発表申込みデータを提供して下さった言語処理学会に深く感謝いたします。また、論文から言語資源名を自動抽出するツールを作成して下さった三菱電機株式会社の谷垣宏一氏に深く感謝いたします。

[3] cor 分野に特有の現象であり、全体の抽出精度は 2.2 節で述べた 92%と大きく変わらないと考えている。