

自然言語処理研究の動向と課題

野村浩郷 田中裕一 徳永建伸 内藤昭三
(九州工業大学) (ICOT) (東京工業大学) (NTT)

1981年に自然言語処理研究会が発足してから10年が経過した。研究活動は極めて活発であり、着実に発展してきた。そこで、この間の研究発表の動向を外国における研究発表とも比較しながら整理し、今後の研究課題の概要を示すことにより、自然言語処理研究会のさらなる発展を期す。

Progress Report of the SIGNLP of IPSJ

Hirosato NOMURA Yuichi TANAKA Takenobu TOKUNAGA Syozo NAITO
(Kyusyu Inst. of Tech.) (ICOT) (Tokyo Inst. of Tech.) (NTT)

Special Interest Group of Natural Language Processing of IPSJ has been continuing its high activity during this one decade. In this report we present the topics of the papers presented at its periodical conferences. We compare these papers with those of papers presented at the conferences held in abroad. We also list a set of key words for the further elaboration in natural language processing research.

1 はじめに

自然言語処理は高度情報処理の基幹をなす最も重要な技術の一つである。従来から多くの研究・開発が世界中で精力的に進められ、その最も大きな成果として、日本語ワードプロセッサや機械翻訳システムなどが実用化された。これにより技術の基盤の一部が明確になった、真に高度な技術を確立するためには、基礎理論、処理方式、システム化、およびシステム活用法のそれぞれにおいて解決されなければならない課題は多い。

自然言語処理研究会は、1981年に、前身の計算言語学研究会(和田弘主査)を発展させる形で発足した。長尾真、吉田将、田中穂積の歴代主査のご指導と会員諸兄のご尽力とにより、自然言語処理研究会は着実に発展してきた。それから10年を経過した今、一つの節目として、過去10年間の研究発表の動向を振り返り、今後の研究課題を概観することにより、自然言語処理研究会のさらなる発展の一助にしたい。

本稿では、まず、自然言語処理研究会の発展状況の概要を示す。次に、自然言語処理研究会において発表された論文の内容の動向の概略を整理する。続いて、外国において発表された自然言語処理に関する論文の内容の動向の概要を整理し、自然言語処理研究会のもののおおざっぱな比較を行なう。最後に、自然言語処理の課題に関するキーワードの概要を列挙する。

2 発表件数の動向

表1に過去10年間の自然言語処理研究会(以下、研究会)の年別発表件数、分野別内訳を示す。研究会発表総数は、88年が開催数が5回ということもあり少なめであるが、全体的に着実に増加している。特にここ数年の伸びは大きく、このままでは次年度からは開催数を増やすか、一回を2日にするなどの対策が必要になっている。発足当時から開催回数、発表件数とも一定以上の水準を保っている点は、この研究会が昔から盛んな研究会であったことをうかがわせる。

分野別に見ると、機械翻訳、意味解析、構文解析が上位に並んでいる。特に、機械翻訳の82、83年における発表件数は多く、一種のブームであったことをうかがわせる。また、構文解析よりも意味解析にカテゴライズされる研究が多い点は意外であった。この点では、後述するACL Annual Meetingとは大きな違いである。最近の傾向としては、

- 談話・文脈、生成関係の研究発表が着実に増えている
- 形態素に関するものが再び増えてきた(歴史は繰り返す?)
- 言語分析に関する発表が少ない

などが目につく。最後の言語分析に関する発表の減少に関連して、この表からはわからないが、言語学サイドからの発表が昔に比べて少なくなっているように感じられる。自然言語処理には計算機以外のさまざまな分野の知識が必要である。この研究会がそういった言語を取り巻くさまざまな分野の研究者たちの交流の場となることも重要であろう。

表2はACL(Association for Computational Linguistics)のAnnual Meeting(以下、ACL)における同様な統計である。ACLの場合、開催は年1回で、査読があるので、発表件数を単純に研究会と比較してもあまり意味がない。ACLの発表件数に関して、84年はCOLINGと合同で開催された点、82年はACL20周年でパネルが3件あり、その分一般発表が少ない点、の2点だけを注意しておく。

分野別に見ると、ACLでは、言語分析、構文解析が多く、意味解析、文法、談話・文脈、生成が続いている。特に、目につくのは研究会で多かった機械翻訳は少ない点である。ヨーロッパ支部の会議(EACL)では機械翻訳も健闘しているが、ACLではほとんどない。この辺は米国の実情を反映していると思われる。言語分析が多くなっているが、これは言語のある側面を理論的に定式化したというものが多く、どの分野にいれてよいか迷ってここに入れたものが少ない。研究会では、大量データを扱ったものが多いが、ACLでは少なく、その点同じ言語分析でも両者の性質は異なっている。その他、研究会との比較で目につくのは、研究会ではほとんどない音声関係の研究がACLでは、多いという点である。ACLの10年間の音声関係の発表は機械翻訳の発表と数の上では同数である。これは、電子情報通信学会に音声関係の研究会があり、情報処理学会にはないことから、音声関係の研究発表はほとんど電子情報通信学会の研究会でおこなわれているものと推察される。しかしながら、電子情報通信学会の音声研究会では「パターンとしての音声」に関するものが主流であり、「言語としての音声」に関する研究はそう多くはないようである。(書き言葉の)言語処理と音声処理との融合に関する研究には魅力的なテーマが数多く残されている。今後、言語を意識した音声に関する研究発表がこの研究会でも増えることを期待したい。

表3はCOLING(International Conference on Computational Linguistics)における同様な統計である。COLINGの場合も、開催が2年に1回で、査読があるなど、やはり研究会と単純には比較できないものの、世界的な傾向をうかがうことはできる。

分野別に見ると、COLING では、構文解析、意味解析、機械翻訳が安定して上位を占めている点は研究会とよく似ている。以下、文法、生成、談話・文脈と続く。機械翻訳に関しては86, 88年あたりにピークがあるが、これは研究会よりも若干後ろにズレており、機械翻訳に対する取り組みの差がみられるようで興味深い。また文法が多いのはACLと共通する点である。対象とする言語はキーワードの項で述べるようになりかなり広く、枠組みとしてはCG (CUG), LFGなどが用いられている。

3 キーワード

キーワードは、発表タイトルから作成したので、必ずしも正確なものではないが、おおよその傾向をつかむには十分であろう。表4は研究会のキーワード、表5はACLのキーワード、表6はCOLINGのキーワードである。それぞれ、上位、30位程度の頻度のものを挙げた。

研究会では分野として用いたものが当然ながら上位に並んでいる。言語に関しては「日本語」、「英語」、「中国語」という頻度順なものも納得のいくところであろう。「Mu プロジェクト」、「Prolog」、「単一化」、「制約」などのキーワードが現れているのは日本の特徴といえるかもしれない。後の3つは比較的新しいキーワードであるが、結構上位にあるのは興味深いことである。また、単純にキーワードとして数えると「構文解析」より「意味解析」の方が数が多いが、「パーザ」も含めるとこの順位は逆転する。「文生成」あるいは「文章生成」というキーワードも結構多いが、これは機械翻訳の生成研究に含まれるものが多く、文(章)生成単独に関する研究はさほど多くない。もっとも数の多い機械翻訳関連のキーワードに関して、まず、翻訳の方向では、「英日翻訳」が「日英翻訳」の約2倍の頻度で現れている。また、翻訳方式に関して、以前は「トランスファ」が多かったが、最近では「中間言語」の方が頻繁に現れるようである。この傾向とともに「多言語間翻訳」というのも最近になって増えてきたキーワードである。

ACLでは、機械翻訳が少ないということを反映して研究会に比べて翻訳に関するものが少なくなっている。“acquisition”, “unification”, “constraint”, “feature”などが上位に現れているのはACLの最近の理論的傾向を表しているといえるかもしれない。また、“approach”, “model”, “theory”などが多いのもACLの性格を反映していると考えることができよう。

COLINGで特徴的なのは、機械翻訳を中心とする処理の対象言語として、日本語が最上位にきていることである。以下、英語、フランス語、ドイツ語、中国語と続く。また、表では“machine readable dictionary”と一括したが、辞書関係のキーワードがかなり多い。また、この表だけからは推移はわからないが、理論的なキーワードが流行の初期のうちから取り込まれる点もACLと同様である。

4 課題と将来の動向

過去10年の研究動向をふまえ、今後の研究課題と動向を分野別にキーワードを羅列することによって予測する。

音声	言語処理との統合による認識率の向上、韻律の理解/生成、実用面では音声タイプライタの実現
形態素	辞書の整備、並列化/ハードウェア化による高速化、構文解析/意味解析との統合
構文	確率などの導入による優先性の実現、しばらくは、unification-basedの解析が主流になるのでは
意味	ATMS, abduction, 非単調推論などの推論手法を用いた意味解析、確率の導入による優先性の実現
談話、文脈	プラン推論、談話構造理論、状況意味論などの実用化
生成	文脈を考慮した談話生成(文の順序、省略、代名詞化)、意図からの生成、生成文の評価手法の確立
辞書	語彙辞書と知識辞書の統合
文法	実用的かつ大規模な日本語文法の開発
対話	ロバストな解析手法の開発、音声/図形などの統合によるマルチメディア化、対話者のモデルの導入
機械翻訳	翻訳精度評価基準の作成、翻訳者支援ワークベンチの開発
言語分析	コーパスの整備、共用化
情報抽出	コーパスからの辞書構築、要約システムの実現
情報検索	キーワード/フルテキスト検索の共用、連想検索の実現
ツール	言語処理ツールの整備、標準化

5 むすび

自然言語処理研究会が発足してから今日までの10年間の研究発表の動向を、外国における自然言語処理関係の研究発表の動向とも対比しながら示した。また、これらの考察をふまえて自然言語処理に関する課題のキーワードを示した。研究は着実に進展しており、必ずしも平坦ではないが自然言語処理の将来への道は希望に満ちている。会員諸氏の今後のより一層のご活躍を祈るとともに、本稿がその一助ともなれば幸いである。

表1 発表件数の動向(自然言語処理研究会)

開催年	81	82	83	84	85	86	87	88	89	90	91	合計	比率
開催数	3	6	6	6	6	6	6	5	5	6	2	57	
発表件数	15	47	41	37	45	41	50	36	58	63	25	458	
分野別内訳													
音声	0	1	0	0	0	0	0	0	0	1	0	2	0.46
形態素	2	3	1	0	2	0	4	2	1	4	2	21	4.85
構文	1	6	8	4	6	5	7	10	11	9	3	70	16.17
意味	3	6	7	6	7	7	12	1	15	14	5	83	19.17
談話, 文脈	0	1	0	0	3	4	1	0	6	8	2	25	0.00
生成	0	1	1	3	0	3	4	1	2	6	3	24	5.54
辞書	3	1	0	1	0	2	0	1	3	2	0	13	3.00
文法	0	0	1	0	2	1	2	4	0	1	1	12	2.77
対話	0	1	0	1	2	1	3	1	1	0	2	12	2.77
機械翻訳	3	14	15	7	12	9	9	6	10	10	4	99	22.86
言語分析	0	3	4	3	5	1	3	6	2	0	0	27	6.24
情報抽出	1	1	1	3	2	3	3	1	4	3	1	23	5.31
情報検索	0	2	0	1	3	1	0	1	1	1	1	11	2.54
ツール	1	2	2	6	1	1	1	0	1	1	1	17	3.93
会議報告	1	5	1	2	0	3	1	2	1	3	0	19	4.39

表2 発表件数の動向(ACL Annual Meeting)

開催年	81	82	83	83*	85	85*	86	87	87*	88	89	89*	90	合計	比率
発表件数	25	16	25	32	40	41	33	34	50	35	34	43	39	447	
分野別内訳															
音声	2	0	2	2	1	2	4	1	3	0	0	1	1	19	25
形態素	0	0	0	1	1	2	3	1	6	1	0	2	1	18	4.03
構文	6	2	6	8	8	8	4	5	6	7	10	5	7	82	18.34
意味	2	2	1	5	7	1	0	2	3	10	9	1	3	46	10.29
談話, 文脈	2	1	2	3	5	4	6	1	1	1	5	3	3	37	8.28
生成	4	4	3	2	1	4	1	3	1	3	1	5	3	35	7.83
辞書	0	0	0	0	5	1	2	3	4	1	4	0	0	20	4.47
文法	0	1	3	2	4	3	5	8	4	5	0	4	7	46	10.29
対話	1	0	2	2	3	2	1	2	1	1	2	2	1	20	4.47
機械翻訳	0	0	0	1	0	4	0	1	5	1	1	5	1	19	4.25
言語分析	8	6	6	5	3	6	6	6	11	4	2	13	10	86	19.24
情報抽出	0	0	0	0	1	2	0	0	2	0	0	1	2	8	1.79
情報検索	0	0	0	1	0	1	0	0	0	1	0	0	0	3	0.67
ツール	0	0	0	0	1	1	1	1	3	0	0	1	0	8	1.79

*が付いたものはEuropean Chapter

表3 発表件数の動向 (COLING)

開催年	80	82	84	86	88	90	合計	比率
発表件数	92	81	103	150	160	131	717	
分野別内訳								
音声	3	2	5	6	8	9	33	4.60
形態素	3	2	2	7	6	6	26	3.63
構文	10	10	20	19	43	23	125	17.43
意味	21	15	12	15	15	22	100	13.95
談話, 文脈	3	5	7	17	11	6	49	6.83
生成	2	3	11	9	12	16	53	7.39
辞書	4	5	5	10	9	2	35	4.88
文法	6	9	5	17	11	19	67	9.34
対話	7	5	10	5	8	0	35	4.88
機械翻訳	8	14	17	27	24	15	105	14.64
言語分析	3	4	7	2	7	3	26	3.63
情報抽出	2	1	0	5	2	6	16	2.23
情報検索	5	3	0	2	0	1	11	1.53
ツール	15	3	2	9	4	3	36	5.02

表4 キーワード一覧 (自然言語処理研究会)

順位	キーワード	頻度	順位	キーワード	頻度
1	日本語	85	18	Mu プロジェクト	9
2	機械翻訳	50	22	制約	8
3	意味解析	28	22	単一化	8
4	英日翻訳	25	22	文章生成	8
5	構文解析	24	22	Prolog	8
6	文生成	19	26	係り受け解析	7
7	英語	18	26	質問応答	7
7	自動抽出	18	26	多言語間翻訳	7
9	パーザ	17	26	中間言語	7
10	意味表現	16	26	中国語	7
10	形態素解析	16	26	文脈情報	7
12	辞書	14	26	並列処理	7
12	対話	14	33	会話	6
14	概念	13	33	概念体系	6
14	格関係	13	33	共起関係	6
14	談話解析	13	33	語彙	6
14	日英翻訳	13	33	支援環境	6
18	トランスファ	9	33	談話構造	6
18	動詞	9	33	統計情報	6
18	名詞句	9	33	物語文章	6

表5 キーワード一覧 (ACL Annual Meeting)

順位	キーワード	頻度	順位	キーワード	頻度
1	parsing(er)	66	19	dialog(ue)	16
2	natural	39	20	lexical	15
3	semantic(s)	38	20	machine	15
4	grammar	37	22	interpreting(ation)	14
5	structure	29	23	discourse	13
6	generation(or,ve)	27	23	morphology(ic,ical)	13
7	processing(or)	24	23	temporal	13
8	system	24	26	acquisition	12
9	analysis(analyser)	22	26	automatic(automated)	12
10	text	20	26	rule	12
11	language	19	26	unification	12
11	syntax(ctic)	19	30	constraint	11
13	approach	18	30	feature	11
13	computation(al)	18	30	represent(ation)	11
13	formal(ism)	18	30	understanding	11
16	English	17	34	linguistic(s)	10
16	description(al)	17	34	theory	10
16	model	17			

表6 キーワード一覧 (COLING)

順位	キーワード	頻度	順位	キーワード	頻度
1	machine translation	103	17	unification	15
2	parsing (-er, etc.)	90	17	French	15
3	semantic(s)	56	19	discourse	14
4	grammar (-tical, etc.)	50	20	interactive (-ion)	13
4	generation (-ator, etc.)	50	20	interface	13
6	Japanese	42	20	database	13
7	linguistic	36	23	dependency	12
8	syntax (-actic, etc.)	30	23	LFG	12
9	machine readable dictionary	28	25	parallel	11
10	English	27	25	phonology	11
11	model	26	25	German	11
11	understanding	26	28	acquisition	10
13	speech	25	28	constraint	10
14	morphology (-ical, etc.)	21	28	disambiguation	10
14	lexical (-ology, etc.)	21	28	interpretation	10
16	logic	17	28	knowledge	10