平成２８年度科学研究費助成事業（特別推進研究）自己評価書
〔追跡評価用〕

平成２８年５月１２日現在

研究代表者 氏　　名	辻井　潤一	**所属研究機関・ 部局・職 （研究期間終了時）**	東京大学・情報理工学系研究科・ 教授
研究課題名	高度言語理解のための意味・知識処理の基盤技術に関する研究		
課 題 番 号	18002007		
研 究 組 織 （研究期間終了時）	研究代表者　　辻井　潤一（東京大学・情報理工学系研究科・教授）		

【補助金交付額】

年度	直接経費
平成１８年度	73,200 千円
平成１９年度	80,700 千円
平成２０年度	77,700 千円
平成２１年度	79,400 千円
平成２２年度	73,100 千円
総　計	384,100 千円

２－１

1.特別推進研究の研究期間終了後、研究代表者自身の研究がどのように発展したか

特別推進研究によってなされた研究が、どのように発展しているか、次の(1)～(4)の項目ごとに具体的かつ明確に記述してください。

(1) 研究の概要

（研究期間終了後における研究の実施状況及び研究の発展過程がわかるような具体的内容を記述してください。）

本研究では、それまでの言語処理分野での常識を打ち破ることを目指して、（１）HPSG という理論に基づいて、実世界のテキスト処理に適用できる**高効率で高耐性な文解析システム**が構築できること、（２）文の構造解析が**意味に基づく深いテキストマイニング**という応用にとっても有用な技術基盤となること、を実際のシステム構築によって示すことであった。この２つの目的を達成するためには、単一テーマの深堀り的な研究（例えば、素性構造の確率モデル、SuperTagging など）にとどまらず、個別の研究成果をシステムとして統合していくことが不可欠であった。したがって、本研究の推進では、**１．**深堀り研究で一流の成果を挙げられる**若手の優秀な研究者**を集めること、**２．**彼らが協働することで**構造的な言語処理と機械学習の成果を統合**していける環境を作ること、**３．**成果を実世界のテキストマイニングに適用してその有効性を示すこと、に留意した。また、日本の研究成果を国際的に認知させる**国際的な研究ネットワーク**の構築にも努力した。これらの方針により、本研究は、個別の成果の寄せ集めでなく**まとまった、国際的にもインパクトのある成果**を挙げることができた。

プロジェクト終了後の 2011 年、私は大学を早期退職しマイクロソフト研究所に移籍したため、本研究の直接的な継続は、本業とは別に、研究所のインターン研究生の指導、および、日本と英国の大学に在籍している共同研究者を通じて行うこととなった。また、本業では、本研究での経験を医療分野のテキストマイニングやウェッブの意味・知識検索で活用することとなった。以下に、その概略を示す。

A. **構造的な言語処理と機械学習の融合に関する研究**：本研究では、英語の HPSG を中心に研究を進めたが、この枠組みではとらえられない問題がのこされた。一つは、**HPSG が本来的に持つ制約の局所性**から、より**大域的な制約が関与する問題**（たとえば、並列句のスコープ決定）が取り扱えないことである。もう一つは、英語においては、単語切り、形態処理、統語処理、意味処理の分離が比較的明確であるのに対して、中国語を典型とするアジア言語では、この段階的処理の区切りが不明瞭であり、これらの処理のより有機的な結合が不可欠となる。この枠組みの２つの欠陥は、**大域と局所、単語切りと意味処理**といった個別モジュールの確率モデルを複数個のモデルを組み合わせる **Joint モデル**とすることで避けることができる。このようなアイデアから、**中国語の文解析や単語切り、英語並列構造を行うための Joint モデル**を開発し、枠組みのさらなる拡張をおこなった。

B. **生命科学分野でのテキストマイニング**：本研究で開発したシステム（MEDIE、FACTA、PathText など）は、構造的な言語処理が深いマイニングに有効であることを示した。しかし、これらのシステムを実際の生命科学の研究に活用できる有用なものとするためには、**生命科学者とのより緊密な共同研究**により、特定分野に固有な意味や知識の問題に取り組む必要がある。また、言語処理のコアの技術にとどまらず、柔軟で使いやすい**インターフェースやワークフロー設計ツール**が不可欠となる。英国マンチェスター大学の NaCTeM、ケンブリッジの EBI などと共同して、これらの課題を解決する研究を進めた。これらの成果は、数多くの論文として論文誌に発表してきた。また、これらの成果は、特別推進研究の成果をより社会的なインパクトにあるものにすると同時に、今後の言語処理研究が解くべき問題、**適応型の汎用的言語ツールとそのワークフロー**の問題を明確にすることになった。

C. **医療分野のテキスト処理**：医療分野におけるカルテや退院サマリなどのテキストは、科学論文とは異なり、文法性が低く電文調のものが多い。ソーシャルメディアでの言語も同様な性質を持つ。このようなタイプのテキストの深い処理には、**特別推進研究で開発した手法を大きく変革**する必要がある。マイクロソフト研究所では、このような文法性の低いテキストの処理に取り組み、いくつかの成果を論文誌に発表した。

D. **適応的な言語処理ツールとワークフローソフトウェア**：言語処理の将来課題としては、特別推進研究で研究したような一般的な手法を、特定の応用タスクでの言語にいかに低コストで適応するか、機械学習の訓練データの作成コストを以下に低減するかにある。特別推進研究で開発した言語処理ツール（例えば、Enju）やワークフローのソフトウェア（例えば、GXP や U-Compare）では、この適応の問題は取り扱えなかった。現在、産業技術総合研究所・人工知能研究センターに移籍した後、この適応的な言語処理の研究に取り組んでいる。

1. 特別推進研究の研究期間終了後、研究代表者自身の研究がどのように発展したか（続き）

(2)論文発表、国際会議等への招待講演における発表など（研究の発展過程でなされた研究成果の発表状況を記述してください。）

1　招待講演(国際会議)

1.Keynote Speech, BioNLP Workshop, (Baltimore USA, 2014)
2.Keynote Speech, Paclic, (Taipei Taiwan, 2013)
3.Keynote Speech, Workshop on Parsing Technology (IWPT), (Nara Japan, 2013)
4.Invited Talk, Meta-Net meeting, (Berlin Germany, 2013)
5.Keynote Speech, Workshop on Health Informatics , (Tsukuba Japan, 2012)
6.Keynote Speech, Workshop on Bio-Text Mining (collocated with Lrec), (Istanbul Turkey, 2012)
7.Invited Talk, NLP and Bio-Medical Informatics, organized by NIH (Washington USA,2012)
8.Keynote Speech, NTICIR, (Tokyo Japan, 2011)
9.Keynote Speech, IWSLT, (San Francisco USA, 2011)

2　論文誌　(30 件)

1.Y Xu, L Chen, J Wei, S Ananiadou, Y Fan, Y Qian, Eric I-C Chang, **J Tsujii**. Bilingual term alignment from comparable corpora in English discharge summary and Chinese discharge summary. BMC Bioinformatics 16: 149 (2015)
2.K Yamamoto, N Inoue, K Inui, Y Arase and **J Tsujii**. Boosting the Efficiency of First-order Abductive Reasoning Using Pre-estimated Relatedness between Predicates. International Journal of Machine Learning and Computing, Vol. 5, No. 2, 114-120, April (2015)
3.S Pyysalo, T Ohta, R Rak, A Rowley, H-W Chun, S-J Jung, S-P Choi, **J Tsujii,** S Ananiadou. Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013, BMC Bioinformatics, Vol.16 (S10), June (2015)
4.P Stenetorp, S Pyysalo, S Ananiadou, **J Tsujii.** Generalising semantic category disambiguation with large lexical resources for fun and profit. J. Biomedical Semantics 5: 26 (2014)
5.T Mu, M Miwa, **J Tsujii**, S Ananiadou. Discovering robust Embeddings in (DIS)Similarity Space for High-Dimensional Linguistic Features. Computational Intelligence 30(2): 285-315 (2014)
6.Y Xu, J Hua, Z Ni, Q Chen, Y Fan, S Ananiadou, I Eric, C Chang, **J Tsujii.**　Anatomical Entity Recognition with a Hierarchical Framework Augmented by External Resources. PloS one 9 (10), e108396 (2014)
7.Y Matsubayashi, N Okazaki, J Tsujii. Generalization of Semantic Roles in Automatic Semantic Role Labeling, Information and Media Technologies 9 (4), 736-77 (2014)
8.HC Cho, N Okazaki, M Miwa, **J Tsujii.** Named entity recognition with multiple segment representations, Information Processing & Management 49 (4), 954-965 (2013)
9.X Sun, Y Zhang, T Matsuzaki, Y Tsuruoka, J Tsujii. Probabilistic Chinese word segmentation with non-local information and stochastic training, Information Processing & Management 49 (3), 626-636 (2013)
10.Y Xu, Y Wang, T Liu, **J Tsujii,** E I-C Chang. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge, Journal of the American Medical Informatics Association 20 (5), 849-858 (2013)
11.Y Xu, Y Wang, T Liu, J Liu, Y Fan, Y Qian, **J Tsujii,** EI Chang. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, Journal of the American Medical Informatics Association 21 (e1), e84-e92 (2013)
12.K Taura, T Matsuzaki, M Miwa, Y Kamoshida, D Yokoyama, N Dun, T Shibata, CS Jun, **J Tsujii**. Design and implementation of GXP make - A workflow system based on make. Future Generation Comp. Syst. 29(2): 662-672 (2013)
13.X Sun, N Okazaki, **J Tsujii**, H Wang. Learning Abbreviations from Chinese and English Terms by Modeling Non-Local Information, ACM Transactions on Asian Language Information Processing (TALIP) 12 (2) (2013)
14.Y Xu, Y Wang, JT Sun, J Zhang, **J Tsujii,** E Chang. Building large collections of Chinese and English medical terms from semi-structured and encyclopedia webs, PloS one 8 (7), e67526 (2013)
15.N Nguyen, JD Kim, M Miwa, T Matsuzaki, **J Tsujii.** Improving protein coreference resolution by simple semantic classification, BMC bioinformatics 13 (1), 304 (2012)
16.S Pyysalo, T Ohta, M Miwa, HC Cho, **J Tsujii,** S. Ananiadou. Event extraction across multiple

levels of biological organization. Bioinformatics 28(18): 575-581 (2012)
17.Y Xu, **J Tsujii**, E I-C Chang. Named entity recognition of follow-up and time information in 20 000 radiology reports. JAMIA 19(5): 792-799 (2012)
18.Y Xu, K Hong, **J Tsujii,** E I-C Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. JAMIA 19(5): 824-83 (2012)
19.Y Xu, J Liu, J Wu, Y Wang, Z Tu, JT Sun, **J Tsujii,** E I- Chang. A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. JAMIA 19(5): 897-905 (2012)
20.Y-Z Zhang, T Matsuzaki, **J Tsujii**. Structure-guided supertagger learning. Natural Language Engineering 18(2): 205-234 (2012)
21.T Mu, J Y Goulermas, **J Tsujii**, S Ananiadou. Proximity-Based Frameworks for Generating Embeddings from Multi-Output Data. IEEE Trans. Pattern Anal. Mach. Intell. 34(11): 2216-2232 (2012)
22.D Andrade, T Matsuzaki, **J Tsujii**. Statistical Extraction and Comparison of Pivot Words for Bilingual Lexicon Extension. ACM Trans. Asian Lang. Inf. Process. 11(2): 6 (2012)
23.JD Kim, N Nguyen, Y Wang, **J Tsujii**, T Takagi, A Yonezawa. The GENIA event and protein co-reference tasks of the BioNLP shared task 2011, BMC bioinformatics 13 (Suppl 11), S1 (2012)
24.S Pyysalo, T Ohta, R Rak, D Sullivan, C Mao, C Wang, B Sobral, **J Tsujii.** Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011, BMC bioinformatics 13 (Suppl 11), S2 (2012)
25.S Kocbek, R Sætre, G Stiglic, JD Kim, I Pernek, Y Tsuruoka, P Kokol, S Ananiadou, **J Tsujii**. AGRA: analysis of gene ranking algorithms. Bioinformatics 27(8): 1185-1186 (2011)
26.Y Tsuruoka, M Miwa, K Hamamoto, **J Tsujii**, S Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics [ISMB/ECCB] 27(13): 111-119 (2011)
27.X Wang, I McKendrick, I Barrett, I Dix, T French, J Tsujii, S Ananiadou. Automatic extraction of angiogenesis bioprocess from text. Bioinformatics 27(19): 2730-2737 (2011)
28.JD Kim, T Ohta, S Pyysalo, Y Kano, **J Tsujii**. Extracting Bio-molecular Events from literature - the BioNLP'09 Shared Task. Computational Intelligence 27(4): 513-540 (2011)
29.S Riedel, R Sætre, HW Chun, T Takagi, **J Tsujii**. Bio-molecular Event Extraction with Markov Logic. Computational Intelligence 27(4): 558-582 (2011)
30.Y Kano, M Miwa, K. B Cohen, L Hunter, S Ananiadou, **J Tsujii**. U-Compare: A modular NLP workflow construction and evaluation system. IBM Journal of Research and Development 55(3): 11 (2011)

3. 主要な国際会議(14 件)

1.H Takamura and **J Tsujii**. Estimating Numerical Attributes by Bringing Together Fragmentary Clues, NAACL, Denver, Colorado (2015)
2.G Kontonatsios, I Korkontzelos, **J Tsujii,** S Ananiadou. Using a Random Forest Classifier to Compile Bilingual Dictionaries of Technical Terms from Comparable Corpora, EACL (2014)
3.H Oiwa, **J Tsujii**. Common Space Embedding of Primal-Dual Relation Semantic Spaces, COLING 2014: 1579-1590 (2014)
4.G Kontonatsios, I Korkontzelos, **J Tsujii**, S Ananiadou. Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora, EMNLP 2014: 1701-1712 (2014)
5.J Hajic, J Tsujii. COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland. ACL 2014, ISBN 978-1-941643-26-6(2014)
6.G Kontonatsios, I Korkontzelos, **J Tsujii,** S Ananiadou. Using a Random Forest Classifier to recognise translations of biomedical terms across languages, ACL (2013)
7.A Hanamoto, T Matsuzaki, **J Tsujii**. Coordination structure analysis using dual decomposition, EACL 2012, 430-438 (2012)
8.X Wu, T Matsuzaki, **J Tsujii.** Akamon: An Open Source Toolkit for Tree/Forest-Based Statistical Machine Translation. ACL, 127-132 (2012)
9.J Hatori, T Matsuzaki, Y Miyao, **J Tsujii**. Incremental Joint Approach to Word Segmentation, POS Tagging, and Dependency Parsing in Chinese. ACL (1) 2012: 1045-1053 (2012)
10.P Stenetorp, S Pyysalo, G Topic, T Ohta, S Ananiadou, **J Tsujii**. brat: a Web-based Tool for NLP-Assisted Text Annotation, EACL 2012: 102-107 (2012)
11.A Hanamoto, T Matsuzaki, J Tsujii. Coordination Structure Analysis using Dual Decomposition, EACL 2012: 430-438 (2012)
12.X Wu, T Matsuzaki, **J Tsujii.** Effective Use of Function Words for Rule Generalization in

Forest-Based Translation. ACL 2011: 22-31 (2011)
13.T Hara, T Matsuzaki, Y Miyao, **J Tsujii.** Exploring Difficulties in Parsing Imperatives and Questions, IJCNLP, 749-757 (2011)
14.J Hatori, T Matsuzaki, Y Miyao, **J Tsujii.** Incremental Joint POS Tagging and Dependency Parsing in Chinese, IJCNLP, 1216-1224 (2011)

1.特別推進研究の研究期間終了後、研究代表者自身の研究がどのように発展したか（続き）

(3)研究費の取得状況（研究代表者として取得したもののみ）

特別推進研究が終了した 2011 年 3 月に東京大学を退職し、同年 5 月よりマイクロソフト研究所に移籍した。マイクロソフト研究所に在籍した 2015 年 3 月までの 4 年間は、当機関が公的研究資金を獲得する立場の機関ではないため、私が代表者となる研究費の獲得はない。

2015 年 5 月より、国立研究法人産業技術総合研究所・人工知能研究センターのセンター長に着任し、当研究センターのセンター長として、下記の研究資金を獲得している。このプロジェクトは、日本での人工知能研究の中核拠点を作ることを目的としたもので、私の特別推進研究の成果を直接引き継ぐものではない。しかし、**適応的なモジュールとそれを組み合わせるためのワークフロー**の重要性という、特別推進研究の過程で得た知見がこのプロジェクトの構成に大きな影響をあたえた。また、特別推進研究とその継続として私が研究してきた言語処理の成果を生かす研究項目を、プロジェクトの中で研究項目として取り上げている。

NEDO　次世代ロボット中核技術開発
次世代人工知能技術分野/人間と相互理解できる次世代人工知能の研究開発
2015－2016 年度　（2 年間）2,063,675 千円
＃プロジェクトは、合計 5 年間の予定

(4)特別推進研究の研究成果を背景に生み出された新たな発見・知見

１．個別言語の持つ普遍性と特殊性
HPSG や CCG のように言語学からの文法理論は、**人間言語がもつ普遍性**、言い換えると、英語、中国語、日本語のような個別言語も、共通の普遍的な枠組みでとらえられることを前提としている。特別推進研究で構築した構文解析のソフトウェアが、大きな変更なく、英語と日本語に適用できたことは、この普遍性を示している。一方で、後続する研究で、中国を取り扱った場合には、HPSG の提供する文法枠組みの普遍性は使えるが、単語切りと構文、意味処理との緊密な連携が必要となり、確率モデルのほうは **Joint モデル**に切り替えざるを得なかった。この確率モデルの強い言語依存性は、現在の技術段階が持つ限界からなのか、より本質的なものなのかを明らかにすることは、適応型の一般的な言語処理モジュールの構築を考える場合、重要な課題となる。

２．分野適応型の言語処理モジュールとワークフロー
生命科学の深いマイニング処理では、構文解析、意味、知識処理の各コンポーネントを特定分野に適応させる必要があった。生命科学といっても、代謝系の研究とシグナル系の研究とでは、取り出したい情報そのものが変化する。活用したテキストのタイプ（実験レポートなど）も変化する。このような適応処理を Joint モデルで行うコストは大きく、**深いマイニング技術のボトルネック**となることが明らかとなってきた。工学的にはコストの高い Joint モデルはむつかしく、Staged Architecture を採用せざるを得ない。現在は、各段階での N-best を組み合わせ爆発を避ける形で出力し、それ以降の処理で Reranking する現実的な枠組みとして研究を進めている。この緩やかな融合モデルが、中国語処理においても可能かどうかなど、今後明らかにすべき問題は多い。また、特定分野への適応を行うための**訓練データのアノテーション・コストを削減するツール群**が不可欠なことなども、多くの研究グループの共通認識となってきている、

３．形態素処理、構文処理、意味処理、知識処理の相互関係
１と２のように、普遍性と特殊性、処理の分野や言語への依存性などは、従来から抽象的には議論されてきた。ただ、このような議論が技術の詳細なレベルで議論できるようになったこと、また、その議論が工学的な応用において重要であることが認識されるようになったことは、今回の特別推進研究である**言語処理にもとづく深いマイニング技術**の有効性が広く認識されてきた結果である。

2.特別推進研究の研究成果が他の研究者により活用された状況

特別推進研究の研究成果が他の研究者に活用された状況について、次の(1)、(2)の項目ごとに具体的かつ明確に記述してください。

(1)学界への貢献の状況（学術研究へのインパクト及び関連領域のその後の動向、関連領域への関わり等）

A. **言語学の理論に基づく構文解析：言語理論に基づく構文解析**は、我々だけでなく、エディンバラ大学、ケンブリッジ大学、ペンシルベニア大学でも取り上げられ、我々が提唱した**コーパスに基づく文法開発、スーパータギング、CFG 近似、素性構造のための確率モデル**が、HPSG、LTAG、CCG という文法枠組みを超えて有効であることが実証的に示されつつあり、グループ間の文法枠組みの差異を超えた標準的な手法として定着してきた。また、素性文法と確率モデルの融合という研究方向は、私自身が中国語の形態素解析、構文解析の Joint モデルでその有効性を示すなど、多様な発展を遂げつつある。これらの発展に我々の研究成果が大きな影響を持ったことは、高い論文引用数に示されている。本研究で開発した**構文解析システム Enju** は、英国マンチェスター大学をはじめとする多くの研究グループで現在も活用されている。また、同じソフトウェアを使った日本語文法の開発も、東京大学、情報学研究所、御茶ノ水大学で行われ、東大入試を目指す AI プロジェクト（通称：東ロボ）の基盤ソフトの一つともなっている。

B. **構文解析のテキストマイニングへの適用**：我々の研究以前のマイニングは、テキストを構造のない単語集合と捉えるものが大半であった。言語処理に基づく深いマイニング技術の重要性、とくにテキストからの事象情報の獲得のための述語―項構造の有効性は広く認識され、**米国の Deep Reading プロジェクト、その後継である Big Mechanism** でも標準的な手法となってきている。本プロジェクトで開発した事象認識プログラム EventMine は、多くのグループから引用され、その後、European Bioinformatics Institute(EBI)や米国 NIH で同様なプログラムの開発が行われることになった。

C. **生命文献を対象としたテキストマイニング**：特別推進研究では、開発した手法の有効性を示すために生命科学分野を対象としたテキストマイニングのシステム（MEDIE,FACTA など）を開発すると同時に、この分野の重要性をアピールするために **ACL のもとに SIG-BioMed** を組織した。私は、現在も、SIG－BioMed の組織委員となっているが、その活動はますます活発になり、言語処理の一大応用分野として認識されるようになった。特別推進の研究チームが中心に組織した SharedTask は、その後、SIG－BioMed の標準タスクとして毎年実施されるようになり、現在も、旧メンバーがその組織の中心的な役割を担っている。この SgaredTask は、**米国 NIH とヨーロッパ EBI** が BioCreative の SharedTask を開始する際のモデルともなった。2015 年の BioCreative には、**私が招待講演者に指名される**ことになった。

D. 科学論文からのテキストマイニングという分野の重要性も、幅広く認識されるようになり、日本では、**2015 年からの CREST プロジェクト（松本奈良先端大教授）**に引き継がれている。また、**米国 DARPA の BigMechanism（2014 年開始）**も、本特別推進の目的意識を引き継いでいる。

E. 特別推進研究の広い意味でのインパクトは、**「大規模なテキストを対象にしても、構造処理や意味処理まで行うことで、深い意味レベルでのマイニングが可能である」**ことを実証したことにある。我々の研究の後、固有表現認識、事象認識など、構造に基づくテキスト処理を大規模テキスト集合に適用することを前提にした意味検索システム、あるいは、その結果を使うマイニングの研究が数多く行われることになった。必ずしも、我々の研究成果だけがこういった傾向を作り出したということではない。ただ、特別推進研究での我々の方向が、この傾向を先取りした先駆的な試みであったことは確かである。

2. 特別推進研究の研究成果が他の研究者により活用された状況（続き）

(2)論文引用状況（上位１０報程度を記述してください。）
表中の引用数は、2016 年 5 月 10 日現在の Google Scholar による。

【研究期間中に発表した論文】

No	論文名・著者名・発行年・ページ数等	日本語による簡潔な内容紹介	引用数
1	Overview of BioNLP'09 shared task on event extraction JD Kim, T Ohta, S Pyysalo, Y Kano, J Tsujii Proceedings of the Workshop on Current Trends in Biomedical Natural Language, 2009	特別推進研究のグループが ACL の SIG－BioNLP のために組織した SharedTask の展望論文。この SharedTask が、その後、米国と EU が中心となった BioCreative での SharedTask のひな形となった。	372
2	Corpus annotation for mining biomedical events from literature JD Kim, T Ohta, J Tsujii BMC bioinformatics 9 (1), 1, 2008	生命科学文献における事象認識のためのアノテーション手法の報告。これ以前には、新聞記事など一般的なドメインの事象認識のみが取り扱われていた。	304
3	Text mining and its potential applications in systems biology S Ananiadou, DB Kell, J Tsujii Trends in biotechnology 24 (12), 571-579, 2006	生命科学者と共著で、生命科学分野でのテキストマイニングの可能性を論じた論文。国際的なインパクトだけでなく、それまでテキストマイニングとは無縁であった生命科学者に分野の重要性を認識させることとなった。	267
4	Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. K Sagae, Jun'ichi Tsujii EMNLP-CoNLL 2007, 1044-1050,2007	文解析に機械学習を取り入れた論文。複数個の文解析プログラムを共同させることで、精度が高くなること、これにより、決定的な解析を行うことで処理効率も上がることを示した。	195
5	Feature forest models for probabilistic HPSG parsing Y Miyao, J Tsujii Computational Linguistics 34 (1), 35-80, 2008	素性構造にもとづく文法のための確率モデルを提唱し、その理論的な正当性を示した。掲載誌は、計算言語の分野で最も権威の高い論文誌。	164
6	Event extraction for systems biology by text mining the literature S Ananiadou, S Pyysalo, J Tsujii, DB Kell Trends in biotechnology 28 (7), 381-390, 2010	生命科学テキストからの事象認識システムについての詳細。これも、生命科学者との共著で、生命科学分野へのインパクトが高い論文となった。	156
7	Evaluating contributions of natural language parsers to protein–protein interaction extraction Y Miyao, K Sagae, R Sætre, T Matsuzaki, J Tsujii Bioinformatics 25 (3), 394-400, 2009	特別推進研究の主目的である理論に基づく文解析が、生命分野のテキストマイニングという応用にとって有効であることを定量的に示した。	142
8	Event extraction with complex event classification using rich features M Miwa, R Sætre, JD Kim, J Tsujii Journal of bioinformatics and computational biology 8 (01), 131-146, 2010	事象認識システム EventMine に関する論文。EventMine は、この後、国際的な SharedTask で常にトップクラスの性能をしめすものとなった。	130
9	Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. HW Chun, Y Tsuruoka, JD Kim, R Shiba, N Nagata, T Hishiki, ... Pacific Symposium on Biocomputing 11, 4-15, 2006	遺伝子と病気の相互関係に注目した事象認識システムの報告。G-D の相互関係の認識をおこなった先駆的な論文。生命科学から医学への橋渡しをする研究として、現在は、多くの研究グループで取り組まれている。	126
10	FACTA: a text search engine for finding associated biomedical concepts Y Tsuruoka, J Tsujii, S Ananiadou Bioinformatics 24 (21), 2559-2560, 2008	２つのエンティティ間の間接的な関係を取り出すシステム。その後、このシステムのための使いやすい Visualizer をマンチェスター大学が開発し、現在も生命科学者によく使われている。	118

【研究期間終了後に発表した論文】

No	論文名	日本語による簡潔な内容紹介	引用数
1	BRAT: a web-based tool for NLP-assisted text annotation P Stenetorp, S Pyysalo, G Topić, T Ohta, S Ananiadou, J Tsujii Proceedings of the Demonstrations at the 13th Conference of the European, 2012	テキスト・アノテーションのためのソフトウェア、優れた視覚化プログラムを備えている。オープンソースとして公開したことにより、国内外の多くのグループで使われている。第一著者は、特別推進研究に博士学生として参加、現在は英国 UCL の研究員	197
2	Overview of BioNLP shared task 2011 JD Kim, S Pyysalo, T Ohta, R Bossy, N Nguyen, J Tsujii Proceedings of the BioNLP Shared Task 2011 Workshop, 1-6, 2011	BioNLP の SharedTask の展望。Bossy 以外は、特別推進に参加した研究員と博士学生。Bossy は、フランス CRNS の研究者。	147
3	Discovering and visualizing indirect associations between biomedical concepts Y Tsuruoka, M Miwa, K Hamamoto, J Tsujii, S Ananiadou Bioinformatics 27 (13), i111-i119, 2011	前頁論文１０の FACTA に、視覚化プログラムとユーザインターフェースを付け加えた論文。このシステムが、現在、英国テキストマイニングセンターでユーザに公開されている。	59
4	The genia event and protein coreference tasks of the BioNLP shared task 2011 JD Kim, N Nguyen, Y Wang, J Tsujii, T Takagi, A Yonezawa BMC bioinformatics 13 (11), 1, 2012	BioNLP の SharedTask の展望。Kim 博士が、DBCLS に移籍したのちに書いた論文。	53
5	Incremental Joint POS Tagging and Dependency Parsing in Chinese. J Hatori, T Matsuzaki, Y Miyao, Jun'ichi Tsujii IJCNLP, 1216-1224, 2011	言語の構造処理を StagedArchitecture ではなく、いくつかの段階を統合した Joint モデルで行った。	45
6	Design and implementation of GXP make—A workflow system based on make K Taura, T Matsuzaki, M Miwa, Y Kamoshida, D Yokoyama, N Dun, ... Future Generation Computer Systems 29 (2), 662-672, 2013	言語の各種のツールを組み合わせたワークフローを分散計算環境で効率よく実行するためのソフトウェアを開発した。特別推進研究で開発されたものをソフトウェアとして整備し、詳細な性能評価も行った。	36
7	Event extraction across multiple levels of biological organization S Pyysalo, T Ohta, M Miwa, HC Cho, J Tsujii, S Ananiadou Bioinformatics 28 (18), i575-i581, 2012	生命科学の分野では、固有名もその内部構造を持つことから、単一の固有名の内部に現れる関係を認識する手法を示した。	33
8	U-Compare: A modular NLP workflow construction and evaluation system Y Kano, M Miwa, KB Cohen, LE Hunter, S Ananiadou, J Tsujii IBM Journal of Research and Development 55 (3), 11: 1-11: 10, 2011	前頁１７の論文にワークフローの実際の使用例とその効率を追加し、IBM が発行する論文誌に発表したもの。ワークフローの基盤には、IBM の UIMA を使った。	32
9	Mining metabolites: extracting the yeast metabolome from the literature C Nobata, PD Dobson, SA Iqbal, P Mendes, J Tsujii, DB Kell, ... Metabolomics 7 (1), 94-101, 2011	同じ化合物が文脈によって酵素として機能したりしなかったりすることから、メタボロームの固有名認識は、一般分野の固有名認識よりもはるかに困難なもとなる。この問題を扱った最初の先駆的な論文。	29
10	Using workflows to explore and optimise named entity recognition for chemistry BK Kolluru, L Hawizy, P Murray-Rust, J Tsujii, S Ananiadou PloS one 6 (5), e20181, 2011	ケンブリッジ大学が開発した化合物名認識プログラムを言語処理ツールのワークフローに組み込むことで、代謝系の固有名認識、事象認識の精度が向上することを示した。	27

3.その他、効果・効用等の評価に関する情報

次の(1)、(2)の項目ごとに、該当する内容について具体的かつ明確に記述してください。

(1)研究成果の社会への還元状況（社会への還元の程度、内容、実用化の有無は問いません。）

１．特別推進研究、および、その後継プロジェクトで作られた意味検索のシステム群（MEDIE、FACTA、PathText など）は、いずれも生命科学分野を対象としたものである。これらは、さらに改良が重ねられ、現在、**英国マンチェスター大学のテキストマイニング・センター（NaCTeM）で実用に供されている。**

MEDIE：http://www.nactem.ac.uk/medie/
FCATA: http://nactem.ac.uk/facta/
PathText: http://nactem.ac.uk/pathtext/

２．英語の**構文解析システム Enju** は東京大学に知財登録され、マンチェスター大学、米国ベンチャー企業にライセンスされた。また、Enju のソフトウェアをもとに作られた日本語の構文解析システムは、東ロボのプロジェクト、東京大学の工学教育プロジェクト、現在進行中の CREST プロジェクト（代表：黒橋・京都大学教授）などで活用されるなど、言語処理の基礎ツールとして普及している。

Enju: http://nactem.ac.uk/enju/index.ja.html、http://kmcs.nii.ac.jp/enju/?lang=ja

３．生命科学分野のテキストマイニングは、特別推進研究のグループ、コロラド大学、コロンビア大学、米国の MITRE の研究者などが始めたものであるが、その後、米国 NIH、EU の EBI、ケンブリッジ大学など、有力な研究グループが参入することとなった。その結果、ACL（計算言語学会）の中にも SIG-BioMeD が組織され、言語処理の応用分野として確立した。また、生命科学の学会（ISMB など）にも、同様な SIG が作られるなど、多くのグループが研究を進めている。この結果、**生命科学や医学の分野の論文の中にもテキストマイニングを道具として使った論文**が出版されるようになっている。必ずしも、我々の特別推進だけがこの分野を作り出したわけではないが、先駆的な研究グループの一つとして、大きな貢献をしたと思っている。

４．特別推進では応用分野を生命科学にとったが、その成果は**医学や医療の分野**にも適用されてきている。私も、マイクロソフト研究所では、中国語の医療カルテを取り扱う研究を行ったが、医学・医療分野でのテキストマイニングの需要は、生命科学に比べるとはるかに大きく、より大きな社会還元が期待できる。米国 DARPA の Big Mechanism は、我々が特別推進研究で開発した PathText と同様な構想に基づいているが、がん研究など、医学研究へと結び付けていることが特徴である。この方向での研究は、**創薬や新たな治療法**など、より直接的で大きな社会貢献が期待できる。

５．言語の意味を活用する応用は、テキストマイニングだけではない。私自身も、マイクロソフト研究所では自然言語で会話を行うシステムの研究に従事してきた。また、**イメージからテキスト**への変換、深い意味を考慮した、**質の高い翻訳システム**など、その応用範囲は極めて広い。特別推進研究で構築した文解析システムが出力する述語―項構造は、このような応用においても、**文の意味へのインターフェース構造**として、その有効性が確認されてきている。今後、我々の成果がこのような多様な応用システムで使われていくことを期待している。私がセンター長を務める人工知能研究センターでも、この方向での研究を開始したところである。

3.その他、効果・効用等の評価に関する情報（続き）

(2)研究計画に関与した若手研究者の成長の状況（助教やポスドク等の研究終了後の動向を記述してください。）

プロジェクトに参加した助教

１．宮尾祐介　情報学研究所　准教授

2012 年に文科省より文部科学大臣表彰若手科学者賞、2015 年に学術振興会より日本学術振興会賞を受賞するなど、言語処理分野のリーダーとして活躍している。

２．松崎拓也　名古屋大学　准教授

情報学研究所の特任助教を経て現職。東大入試に合格する AI システム（東ロボ）で、言語理論に基づく論理意味論を使った研究で影響力のある論文を発表するなど、活発な研究活動を行っている。

３．鶴岡慶雅　東京大学　准教授

プロジェクト当時、英国マンチェスター大学研究員であったが、私（辻井）がマンチェスター大学教授を兼任していたことから、本プロジェクトに参加。マンチェスター大学の後、北陸先端大学院を経て現職。機械学習と言語処理とを統合する研究、コンピュータ・ゲーム（将棋）の研究で活躍。

プロジェクト雇用の研究員

１．JinDong　Kim　　情報システム研究機構　DBCLS　准教授

生命科学分野でのテキストマイニングの国際的なリーダーとして活躍。プロジェクト終了後、BioNLP の SharedTasks のオーガナイザー、国際論文の特集号のエディターを務める。

２．Kenji　Sagae　　米国 USC　ISI, Assistant Prof.　(8 月より UCDavis　Associate Prof.)

米国での言語処理研究の中核の一つである USC の ISI で研究をすすめ、8 月より UC Davis のテニュアトラックの Assistant Prof として採用されることが決まっている。

３．Rune Saetre　　ノルウェイ・NTNU　Associate Prof.

NTNU(Norwegian University of Science and Technology)の人工知能分野の准教授として、医療分野のテキストマイニングの研究に従事している。

４．岡崎直観　　東北大学　准教授

2016 年に学術振興会から日本学術振興会賞を受賞するなど、言語処理分野のリーダーとして活躍。

５．狩野芳伸　　静岡大学　准教授

特別推進研究において言語処理ツール群の開発を行ったのち、その成果をさきがけ（JST）で発展させた。その研究は、インターオペラブルな言語処理ツールの開発を目指した EU プロジェクトに影響を与えている。

６．三輪誠　　豊田工業大学　准教授

生命科学分野での事象認識システム（EventMine）を開発、EventMine は現在も英国マンチェスター大学で運用され、生命科学以外の様々な分野での事象認識に活用されている。

７．Sampo Pyysalo　　英国ケンブリッジ大学　研究員

JinDong Kim と同様に、生命科学分野でのテキストマイニングの国際的なリーダーとして活躍。ケンブリッジ大学では、テキストマイニングの対象分野を生命科学から医療分野へと拡張する研究に従事。

８　Sebastian　Riedel 英国 UCL　Reader

プロジェクト終了後、米国マサチューセッツ大学の研究員を経て現職。マルコフ・ロジック・ネットワーク（MLN）のソフトウェアを公開するなど、機械学習を言語処理に適用する分野の国際的なリーダーとして活躍。Reader は、英国の大学では若手教授のクラスに対応。

９．綱川隆　　静岡大学　助教

ウェッブのための言語処理、専門用語の自動抽出などの研究を行っている。

１０．原忠義　　東京大学　特任研究員

情報学研究所の特任助教を経て、現職。特別推進研究の成果である言語の構造情報をテキストマイニングに反映する研究を特許文献に対して適用する研究を行っている。

１１．増田勝也　　東京大学　特任助教

カリキュラム、シラバスなど、教育に関するテキストを対象にしたテキストマイニングの研究に従事。