# 機械翻訳向け自動コーパス生成

Automatic Corpora Generation for Machine Translation

山内真樹 Masaki Yamauchi
藤原菜々美 Nanami Fujiwara
今出昌宏 Masahiro Imade

**要　旨**

機械翻訳（MT：Machine Translation）は，大量の対訳コーパス（原言語と目的言語の文章対データ）から，翻訳に必要なモデルを獲得することで翻訳を行う．対訳コーパス数の衆寡が翻訳の性能に直結する一方で，対訳コーパスの収集・獲得は高コストであり，MTの実用化課題である．筆者らは「少量の対訳コーパスからコーパス候補文を生成し，識別器により選択」する自動コーパス生成技術を開発している．翻訳性能を対コーパス数比で換算した場合，従来手法が約1.5倍程度の生成効率であるのに対し，本手法では約12倍以上の文生成を達成した．客観評価（BLEUスコア）でも約2.5ポイント向上し顕著な効果を有している．コーパス生成の自動化により，生成コストの削減と翻訳性能の向上とを両立した．

**Abstract**

Recently, Machine Translation (MT) systems have become widely used. MT is a system whose model is trained on large quantities of parallel corpora. Although the size of the corpora directly affects the performance of MT, corpora data is expensive in general, and hence an effective method is needed to create parallel corpora. From this perspective, we have been developing a unique technique to automatically generate large-size corpora from a small number. The method consists of 1) creating candidate sentences by utilizing various expressions and paraphrases on the Web, and 2) choosing correct sentences with a machine learning verifier. In this letter, we report that MT performance using the corpora generated by our system is improved more than 2.5 points compared to the original small amount of corpora. This method will dramatically improve translation quality while reducing the cost for corpora preparation.

## 1. はじめに

筆者らは，機械翻訳（MT：Machine Translation）向けの自動対訳コーパス生成（ACG：Automatic Corpora Generation）技術を開発した．

MTの構築には約100000文～1000000文単位での対訳コーパス（学習用の翻訳文データ）が必要であるが，対訳コーパスの収集は一般に高コストである[1]．特に初期段階で準備できる対訳コーパス量は，経験的に1000～10000文程度である．少量の対訳コーパスでは学習に十分な情報が得られず，MTの性能は著しく低下する．得られる文量と必要な文量との差は10倍以上にのぼるが，従来手法での対訳コーパス生成手法ではおよそ1.2倍から1.5倍程度[2], [3]であり，初期段階で準備できる少量コーパスからのMT構築は極めて挑戦的な課題であった．

これに対し筆者らの開発したACG法は，言い換え表現に着目し類似候補文の生成と識別という新しい枠組みを導入することで，10倍以上のコーパス自動生成を達成した．さらに類似候補文の識別に関して，利用者およびシステム改善協力者からのフィードバックを教師データとする仕組みを併せて導入することにより，漸進的に識別器の性能向上を図り，翻訳性能の向上につなげている．これらにより，実用化時にはコーパス生成コストの削減を期待できる．

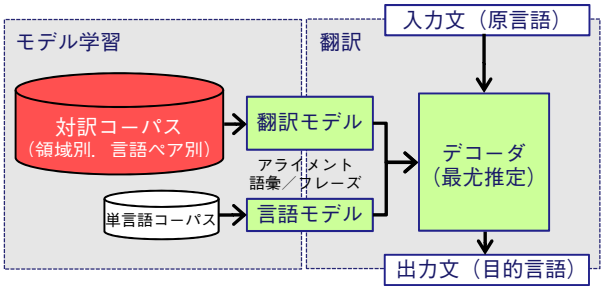
## 2. 機械翻訳（MT）

本章ではMTについての概要を説明する．MTは大きく2つの種類に分けられる．1つは統計的機械翻訳（SMT: Statistical MT），もう1つはニューラルネットワーク型機械翻訳（NMT：Neural network MT）である．いずれの手法においても，大量の対訳コーパスを基にして翻訳に必要なモデルの学習を行う．両者の違いは学習および翻訳に用いる手法が，翻訳対象となる両言語間の特徴量を「明示的にフレーズの統計的な確率分布に基づかせる」（SMT）か，「非明示的にニューラルネットワークに直接獲得させるか」（NMT）にある[4], [5].

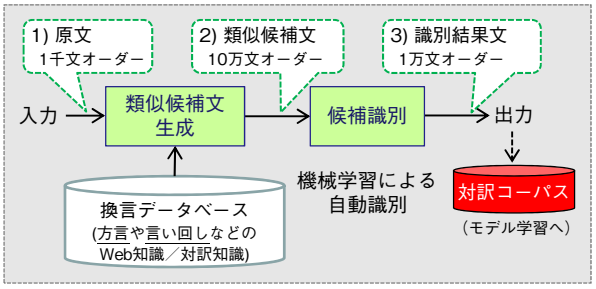
統計的機械翻訳（SMT）は，大きく2つの要素から構成される．1つは事前に統計的な翻訳モデル・言語モデルを構築する「モデル学習」，もう1つは事前に構築されたモデルに従い，最尤（さいゆう）推定により入力文（原言語）から確率的に最適と推定される訳文を出力文（目的言語）として出力する「翻訳エンジン」である．

SMTの構成概略図を**第1図**に示す．

「モデル学習」では，対訳コーパス・単言語データを用いて統計的に翻訳モデル・言語モデルを構築する．原言語文を$J$，目的言語文を$E$とすると，原言語から目的言語への翻訳は確率$P(E|J)$の最大化タスクとなり，ベイズの定理から次の（1）式となる．




第1図　統計的機械翻訳の概略図
Fig. 1　Illustration of statistical machine translation




第2図　自動コーパス生成技術の構成概要図
Fig. 2　Illustration of automatic corpora generation

$$P(E\mid J)=\frac{P(J\mid E)P(E)}{P(J)}\propto P(J\mid E)\quad\cdots\cdots\cdots\quad(1)$$

言語モデルは$P(E)$に相当し，目的言語らしさ（流暢（りゅうちょう）さ）に寄与する．翻訳モデルは$P(J|E)$に相当する．“ある目的言語文（単語／句）が，ある原言語文（単語／句）であった確率”で，対訳コーパスに含まれる各言語・全単語での共起確率などを用いて統計的に獲得される[6].

翻訳エンジンは，翻訳モデルおよび言語モデルをもとに，入力文に対して確率が最も高い目的文候補を組み合わせ問題として決定する．統計的に得られた確率分布をもとに推定を行うため，SMTの翻訳性能は対訳コーパスの質と量に大きく依存する[7].

ニューラルネットワーク型機械翻訳（NMT：Neural network MT）では，翻訳モデルがニューラルネットの重みとして学習されるため[5]，翻訳性能はニューラルネットの各種パラメータや学習回数にも依存する．しかし学習データ，すなわち対訳コーパスが最も重要であることは言をまたない．

## 3. 自動対訳コーパス生成（ACG）

筆者らが開発している自動対訳コーパス生成（ACG）技術の構成概要図を**第2図**に示す．ACGは，対訳コーパス入力から「類似候補文生成」器と「類似候補文識別」器により識別結果文としてコーパスを生成する．

これにより，自動生成する対訳コーパス数となり得る候補文の数を飛躍的に増大させつつ，翻訳機の学習に有効なコーパスのみの抽出が可能となった．

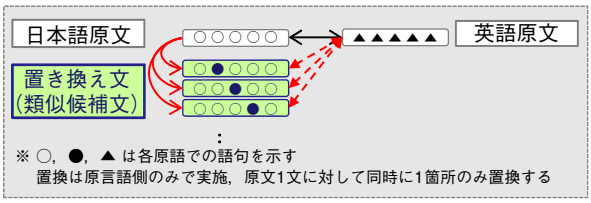
類する先行研究としては，WordNet[8]から言い換え候補を選択し，対訳コーパスの拡張を行う手法[2]や，置き換えルールでのコーパス拡張手法[3]などが挙げられる．

しかし，本稿で述べるような「機械翻訳の学習データとなる対訳コーパスについて，対訳コーパスの類似候補文を生成し，機械学習により候補文の識別を行う」枠組みは，これまでのところ報告例を見いだせておらず，新規性の高いアプローチであると考えられる．

### 3.1　類似候補文生成

「類似候補文生成」器では，言い換え表現を言語資源（WordNet [9]，PPDB（Paraphrase Database）[10]，内容語換言辞書[11]）などを参考として独自に構築し（換言データベース，DB），入力文に対して適用することで類似候補文を得る．

類似候補文の生成模式図を**第3図**に示す．元となる原文内の語句について，換言DBを用いて言い換えることで類似候補文を生成する．ここでは，言い換えを同時に1箇所のみに適用している．同一文中の複数箇所に対して同時に言い換えを行った場合，文意が大きく変化し単言語としては正確であっても対訳コーパスとしては成立しない（両言語間で文意が異なる）ケースを考慮し，言い換えを同時に1箇所のみに適用している．語句の置き換えのみに注目しているため，同時に1箇所までの制限を加えたとしても，類似候補文の中には文全体として見た場合に，不適切な文や破綻した文が生成される可能性がある．要因としては，原文が含まれる対訳コーパスのドメインが換言データベースのエントリと必ずしも合致しないことや，データベース自身のノイズなどに起因する．




第3図　類似候補文の生成模式図
Fig. 3　Paraphrasing illustration

### 3.2　類似候補文識別

「候補識別」器では，類似候補文に含まれる破綻文を

極力排除し，対訳コーパスとして適切な文を選択抽出することを目的としている．識別された文（識別結果文）を新たにMTの訓練文（対訳コーパス）として用いる．

類似候補文から人手で良質な対訳コーパスを抽出した場合，抽出したコーパスから学習した翻訳モデルにおける翻訳性能の向上を確認している[12]．そのため，いかに「類似候補文を自動識別して“良い文”の集合を得られるか」が，ここでの鍵となる．

具体的な識別器の素性としてN-gram[13]を用いる．特に置き換えが発生した箇所に着目し，置き換え箇所を含んだ語句の素性から，“良い文” “悪い文”の識別を行う．

しかしながら，識別に必要な教師データ，特に“悪い文”に相当する教師データは存在しないため，何らかの方法で識別対象となる文を機械識別可能な形で表す必要がある．ここでは，教師なし学習に準じて文の自然さをN-gram（言語モデル）から最尤推定する．具体的には，N-gramの出現頻度を素性として，（1）式および（2）式から各類似候補文の自然さを対数尤度として定義・推定し，閾（しきい）値により自動識別している．

識別にあたっては，まず，置き換えが行われた語句を最低1語含むN-gramを$k$個取得し，それぞれの出現頻度を求める．出現頻度から推定した対数尤度の平均値をその文のスコアとして閾値判定を行う．置き換え対象周辺のフレーズの出現頻度が高ければ，“良い文”として選択される可能性も高くなる[13]．

N-gramの尤度は，単語$\omega_1\omega_2\cdots\omega_n$の尤度を$P(\omega_1\omega_2\cdots\omega_n)$として，以下の式で推定する．

$$P(\omega_1\omega_2\ldots\omega_n)\cong\prod_{i=1}^{n}P(\omega_i\mid\omega_{i-N+1}\ldots\omega_{i-1})\cdots\cdots\quad(1)$$

一般に単語長が長くなればなるほど，出現頻度は低下する傾向があることから，尤度の推定には未知語（未知フレーズ）の影響を考慮する必要がある．ここでは，加算スムージングにより，出現頻度を求める際に，各N-gramの出現頻度に単語列の異なり総数に応じた値を加えることで未知語（未知フレーズ）に対応している．

一般に単語長が長くなればなるほど，出現頻度は低下する傾向があることから，尤度の推定には未知語（未知フレーズ）の影響を考慮する必要がある．ここでは，加算スムージングにより，出現頻度を求める際に，各N-gramの出現頻度に単語列の異なり総数に応じた値を加えることで未知語（未知フレーズ）に対応している．

$$P(\omega_i\mid\omega_{i-N+1}\ldots\omega_{i-1})=\frac{C\left(\omega_{i-N+1}^{i}\right)+\delta}{C\left(\omega_{i-N+1}^{i-1}\right)+\delta V}\quad\cdots\cdots\quad(2)$$

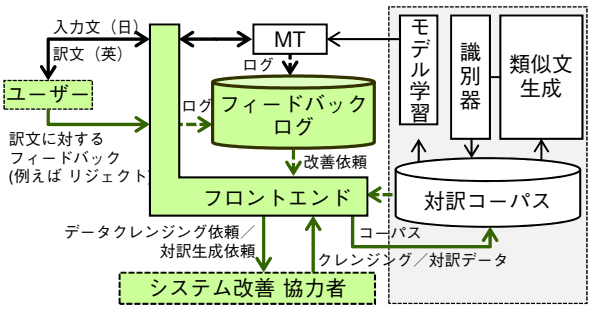
ここで$C(\omega_1^n)$は単語列$\omega_1\omega_2\cdots\omega_n$がN-gram中に出現する頻度，$V$は単語列の異なり総数，$\delta$は定数である．

## 4. ユーザーフィードバックからの学習

本稿でのフィードバックとは，機械翻訳の利用者が翻訳を行った際に，得られた訳文に対して行う品質評価や，それに基づいて類似候補文や訳文の選択・修正を行う人的な作業を意味している．前節までで自動生成したコーパスに対し利用者からのフィードバックを得ることで，さらに漸進的に識別性能を向上させ，結果として統計的機械翻訳やニューラルネット型機械翻訳における翻訳モデル学習の効率化を狙う．これにより，翻訳モデルの学習アルゴリズムを問わず，対訳コーパスから学習を行う機械翻訳機における翻訳性能向上を図る．

利用ログやフィードバックをもとに自律学習できる技術の構築は，翻訳に限らず機械学習で求められる機能の1つである．一方で，翻訳機の学習には本質的に原言語（例：日本語）と目的言語（例：英語）双方の知識が必要となる．そのため，利用ログやフィードバックが学習に有意なデータであるためには，その情報も原言語と目的言語双方に跨（またが）っている必要があった．翻訳機の中心的な利用目的を考慮すると，目的言語側の知識を強く期待することは必ずしも適切ではないため，これまでは活用できるフィードバック情報や取得する仕組みには限界があった．

そこで，ユーザーフィードバックからの学習として，システム改善への協力者について，彼らのもつ言語知識が原言語側（利用者の母国語）のみであっても有効に学習できる枠組みを新たに構築した．概要図を**第4図**に示す[14]．




第4図　フィードバックシステムの概要図
Fig. 4　Illustration of our feedback system

具体的には，「翻訳機のユーザーと，翻訳機のシステム改善協力者とを区別」し，「ユーザー」は，「目的言語側の知識がなくとも，実利用を通じて機械翻訳された対訳文が適切であったかどうかについて知り得る状況」にあり，「訳文の品質を何らかの形で評価できる」ものとする．

また「翻訳機のシステム改善協力者」についても，「特に目的言語側の知識有無を問われない」ものとする．

ユーザーが翻訳器からの出力文（訳文）の品質を低いとした場合，システム改善協力者に対して，識別結果文を含む対訳コーパスから元の入力文に近い文（原言語）を提示する．システム改善協力者は，そのなかから比較的品質の高いコーパス（原言語文として意味をなしているコーパス）を選択する．

選択された比較的品質が高いとされる原言語文について，含まれる語句のN-gram出現頻度を増加させる．これにより，3.2節で述べた類似候補文識別において同じ語句を有する類似候補文が，識別結果文として識別される確率を上げる．

## 5. 実験評価

類似候補文に対し，N-gramを素性として学習した識別器により得た識別結果文を学習用の対訳コーパスとして翻訳機の訓練を行う．訓練より得られた翻訳モデルを用いて別途用意した評価文の翻訳を行い，その訳文品質を客観評価することにより，類似候補文および識別結果文の翻訳性能への効果を確認する．

### 5.1 翻訳結果の評価尺度

MTの翻訳精度を客観的に評価する手法として，あらかじめ準備した参照訳（正解訳）と，MTの出力文を比較する手法が一般的である．このような自動評価手法のうち，最も一般的なものにBLEU（BiLingual Evaluation Understudy）がある[15]．BLEUは，参照訳と出力文との単語列での一致度に基づいて算出され，値が大きいほど適合率が高く，翻訳精度が良いとされる．本稿でもこのBLEUを客観評価指標として用いる．

### 5.2 評価条件

評価対象として，原文規模の異なる2種類のコーパスを準備した．

(A) 旅行／道案内コーパス：
　道案内における行動指示などで使われる言い回しを含んだ対訳コーパス

(B) 医療コーパス：
　医院内での問診などで使われる言い回しを含んだ対訳コーパス

また，フィードバック効果の評価用として，

(C) フィードバック評価用コーパス：
　(A)よりランダムに選択した100文のコーパス

を準備した．

原文(A)，(B)に対し類似候補文・識別結果文を得て対訳コーパスとし，それを用いて翻訳機の訓練を行った．評価文はそれぞれ原文から約300文を抽出して用いた．なお，評価文は訓練文から削除して学習を行っている．使用する各コーパスの文数を**第1表**に示す．

第1表　学習に用いたコーパス文数
Table 1　Number of training corpora

| | (A) 旅行／道案内コーパス | (B) 医療コーパス | (C) フィードバック評価コーパス |
|---|---|---|---|
| (1) 原文 | 5000 | 102 000 | 100 |
| (2) 類似候補文 | 130 000 | 428 000 | 8000 |
| (3) 識別結果文 | 6200 | 335 000 | 100～4400* |

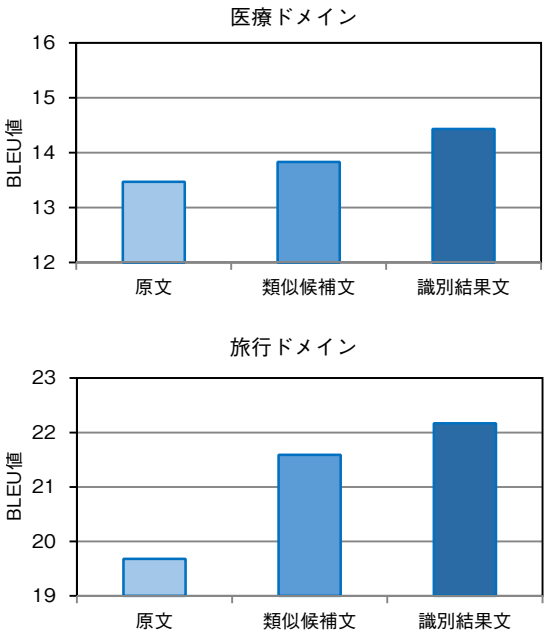
*：フィードバック数に応じて変動

### 5.3 客観評価

**第5図**に(1)原文，(2)類似候補文，(3)識別結果文を用いた翻訳モデルの性能を示す．各翻訳モデルは10回ずつ独立して訓練を行い，その平均BLEU値を示している．

旅行ドメインでは，原文約5000文に対して約12倍の62000文が識別結果文として得られており，また，それを用いた翻訳機の性能も，BLEU値が約2.5ポイント向上と著しい改善が見られた．

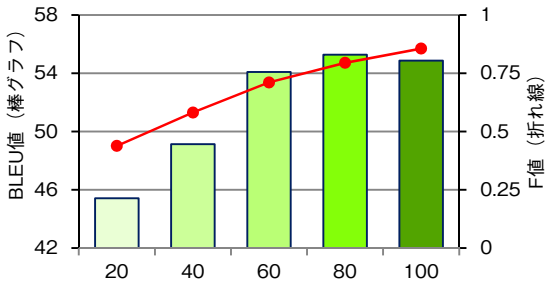
比較的原文数の多い医療ドメインにおいても，約3.3倍のコーパスを自動生成できており，翻訳性能も約1.0ポイントの向上とこちらも大きな改善を確認できた．医療




第5図　自動コーパス生成技術による翻訳性能比較
Fig. 5　Performance evaluation of proposal method

ドメインと比較し，旅行ドメインでは比較的似通った文意や表現が多いことから，より大きな改善結果が得られていると推測される．

また，**第6図**にフィードバックによる識別性能と翻訳性能の変化を示す．(C)の原文100文を入力文として想定し，入力文に近い類似候補文のなかから比較的品質の高いコーパスを選択し，文スコアへの反映を行った．**第6図**中，横軸はフィードバック対象となった入力文数を示し，縦軸はBLEU値および識別器のF値を示している．フィードバックの文数が増えるに従い，翻訳性能および識別器のF値がともに上昇傾向にある．これにより品質が良い類似候補文が適切に識別され，かつ翻訳性能に寄与していることが確認できた．




第6図　フィードバックによる識別性能と翻訳性能
Fig. 6　Performance of our verifier and translator with feedback

### 5.4　多言語翻訳システムへの応用

多言語翻訳試作機のB2B顧客様先でのフィールドテストとして，類似候補文の実証実験導入から応用展開に着手している．システムへの応用にあたっては，特に，”自然さ”スコアの推定処理に時間が掛かっており，識別器の処理時間にも配慮が必要である．識別器の性能を落とさず高速化を行うため，置き換え箇所に限定した注視モデルと識別用データベースのサブセット化を行うことで，識別器を含めたシステム全体についても早期の応用展開を図っている．

## 6. まとめ

少量対訳コーパスの自動生成方法を構築した．機械翻訳の学習に必要なコーパスデータを効率的に生成できるだけではなく，翻訳性能の向上も同時に実現した．

原理的に対象となるドメインや言語を問わないため，非常に汎用性の高いデータ生成技術としての共通技術化が見込まれ，応用範囲も翻訳を前提とした対訳コーパスに留（とど）まらず，言語を用いた対話インタラクション向けのデータ全般に展開が可能である．

本研究にあたり，国立研究開発法人 情報通信研究機構（NICT） ユニバーサルコミュニケーション研究所 隅田副所長および内山研究マネージャーより賜りましたご指導ご助言に感謝致します．

### 参考文献

[1] 内山将夫, ”対訳データの効率的な構築方法,” 情報通信研究機構季報, vol. 58, nos. 3/4, pp.37-43, 2012.
[2] Madnani N. et al., “Generating targeted paraphrases for improved translation,” ACM Trans. Intell. Syst. Technol.4, 3, Article 40, 2013.
[3] Yuval M et al., “Distributional Phrasal Paraphrase Generation for Statistical Machine Translation,” ACM Trans. Intell. Syst. Technol.4, 3, Article 39, 2013.
[4] Philipp Koehn et al., “Statistical Phrase-Based Translation,” NAACL '03 Proc. on Human Language Technology, vol. 1, pp.48-54, June 2003.
[5] Mikel L. Forcada et al., “Recursive hetero-associative memories for translation,” Neuroscience to Technology, vol. 1240, pp. 453-462, June 1997.
[6] Philipp Koehn et al., “Moses: Open Source Toolkit for Statistical Machine Translation,” Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, June 2007.
[7] 松田繁樹 他, “多言語音声翻訳システム”VoiceTra”の構築と実運用による大規模実証実験,” 信学D, no.10, pp.2549-2561, 2013.
[8] George A. Miller, “WordNet: A Lexical Database for English,” Communications of the ACM, vol. 38, no.11, pp.39-41, 1995.
[9] Hitoshi I. et al., “Development of Japanese WordNet,” In LREC-2008, Marrakech, May 2008.
[10] Mizukami M et al., “Building a Free, General-Domain Paraphrase Database for Japanese,” The 17th Oriental COCOSDA Conference, Phuket, Sep.2014.
[11] 山形祐輝 他, “普通名詞換言辞書の構築,” 言語処理学会第20回年次大会発表論文集, pp.7-10, 2014.
[12] 藤原菜々美 他, “自動コーパス生成による少量対訳コーパスからの統計的機械翻訳,” 言語処理学会第22回年次大会論文集, pp.219-222, 2016.
[13] 矢田晋, “N-gram コーパス -日本語ウェブコーパス2010,” http ://s-yata.jp/corpus/nwc2010/ngrams/, 参照 Apr.19, 2017.
[14] 山内真樹 他, “自動コーパス生成とフィードバックによる少量対訳コーパスからの統計機械翻訳,” 2016年度人工知能学会大会, 北九州, June 2016.
[15] Papineni, K. et al., “BLEU: a method for automatic evaluation of machine translation,” ACL: 40th Annual meeting of the Association for Computational Linguistics. PhiladelPhia, July 2002.

## 執筆者紹介

山内　真樹　　Masaki Yamauchi
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
（2017年3月まで先端研究本部に所属）

藤原　菜々美　　Nanami Fujiwara
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
（2017年3月まで先端研究本部に所属）

今出　昌宏　　Masahiro Imade
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
（2017年3月まで先端研究本部に所属）