アダプテーションによる製薬専門NMTの研究開発

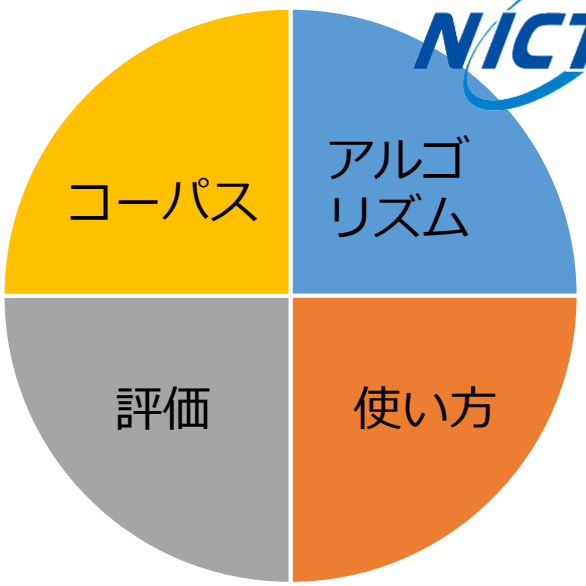
情報通信研究機構
内山将夫

JTF関西セミナー
２０２０年6月２３日

概要

- NMTの概要とみんなの自動翻訳@TexTra
- 製薬専門NMT

コーパスベースＭＴの構成要素



- 自動翻訳アリゴリズム

１９８０年代からの発展⇒ＮＭＴ

- 学習データとしての対訳コーパス

異なる言語で同一意味の文章の組からなるデータベース

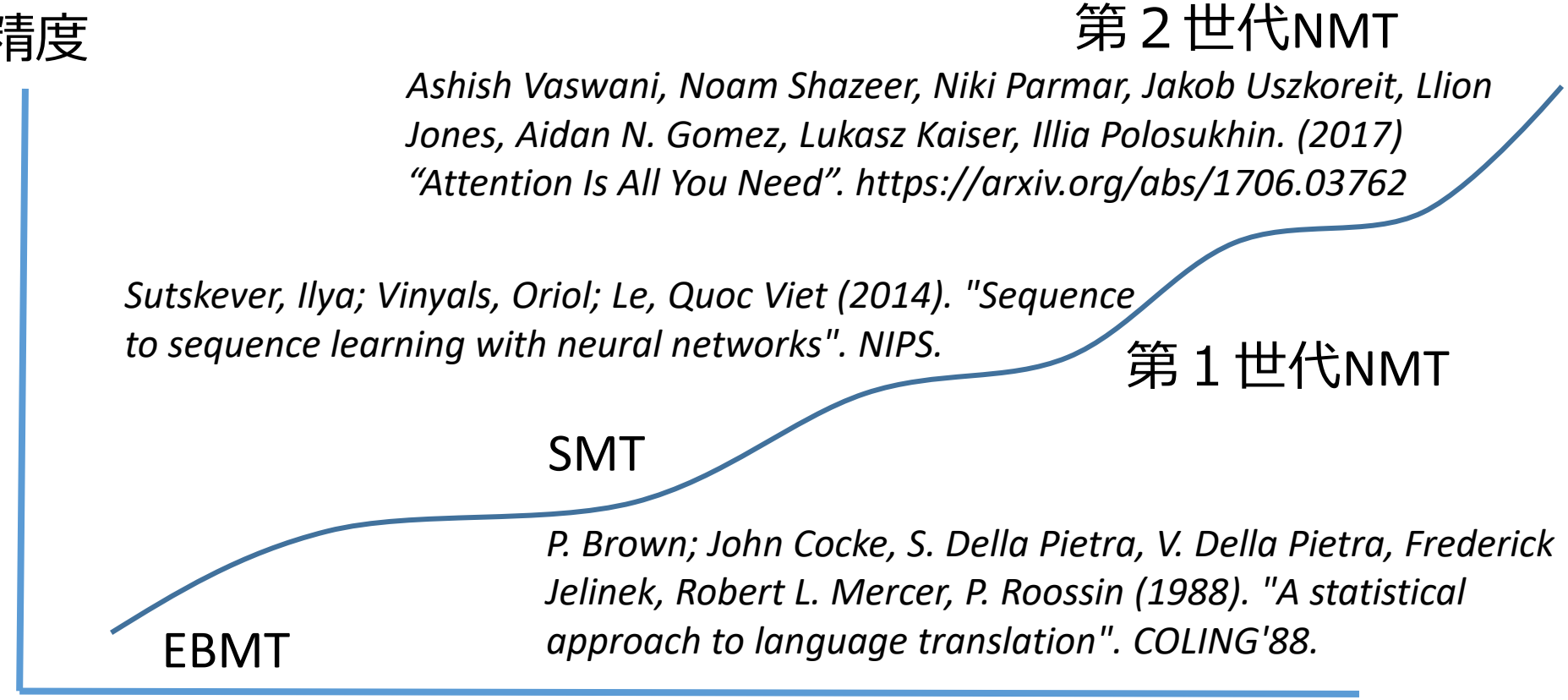
- ＭＴの評価手法

人間がどのように感じるかの観点からのＭＴの良さ

- ＭＴの使い方

ＭＴの普及により、ＭＴの適切な使い方の周知が必要

コーパスベースＭＴアルゴリズムの進展



Makoto Nagao (1984). "A framework of a mechanical translation between Japanese and English by analogy principle". In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers

第２世代NMT（汎用ＮＴ）@みんなの自動翻訳

汎用NT @みんなの自動翻訳

原文英語	汎用ＮＴ（第２世代）
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train.	支配的なシーケンス変換モデルは、エンコーダ/デコーダ構成における複雑なリカレントまたは畳み込みニューラルネットワークに基づいている。 また、最適なモデルは、アテンション機構を介してエンコーダとデコーダを接続する。 本稿では,注意機構のみに基づき,リカレントと畳み込みを完全に省く,新しい簡単なネットワークアーキテクチャTransformerを提案する。 2つの機械翻訳タスクに関する実験では、これらのモデルは、より並列化可能であり、トレーニングに要する時間が大幅に短縮される一方で、品質が優れていることが示されている。

English sentences are part of Abstract from Ashish Vaswani, et al. (2017) “Attention Is All You Need”. https://arxiv.org/abs/1706.03762

汎用NT @みんなの自動翻訳

原文日本語	汎用ＮＴ（第２世代）
2016年のニューラル機械翻訳（NMT）の実用化は、翻訳業界に衝撃を与え、ポケトークのような自動翻訳端末の市場拡大につながるなど、社会に大きなインパクトを与えた。 ただし、翻訳技術や自然言語処理技術（NLP）分野では、その後も革命級のブレークスルーが相次いでいる。 翻訳を含む言語系の人工知能（AI）が従来の常識を次々と塗り替え、ありえないペースで発展している。	The practical application of Neural Machine Translation (NMT) in 2016 gave a shock to the translation industry and had a large impact on society, such as the expansion of the market for automatic translation terminals such as PokeTalk. However, breakthroughs in translation and natural language processing (NLP) have continued. Artificial intelligence (AI), a linguistic system that includes translation, is evolving at an incredible pace, breaking conventional wisdom.

原文は次の一部：野澤 哲生「AI翻訳が人間超え、言葉の壁崩壊へ」日経エレクトロニクス、2019/08/20
https://tech.nikkeibp.co.jp/atcl/nxt/mag/ne/18/00046/00002/

自動評価尺度ＢＬＥＵ

- ＭＴ訳と参照訳の類似度の尺度
- １０００文～５０００文程度の文章で計算
- 一般的にＢＬＥＵ３０～４０以上であれば高精度

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. (2002)
BLEU: a Method for Automatic Evaluation of Machine Translation. ACL

みんなの自動翻訳でのＢＬＥＵ計算法

- 自動翻訳＞自動翻訳評価＞「＋新規登録」

第２世代製薬NMTは高精度

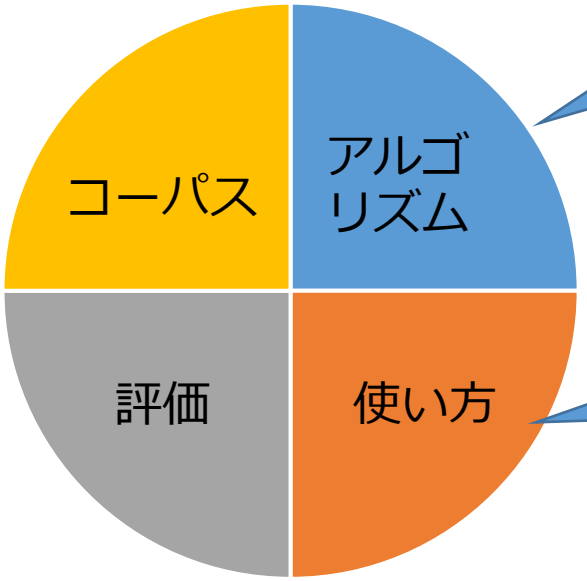
	第１世代	第２世代
汎用	２３．８	３０．３
製薬	３９．９	４９．４

第１世代＜＜第２世代

汎用　＜＜　製薬

自動翻訳尺度BLEUの比較

NMTにより新しくなった点



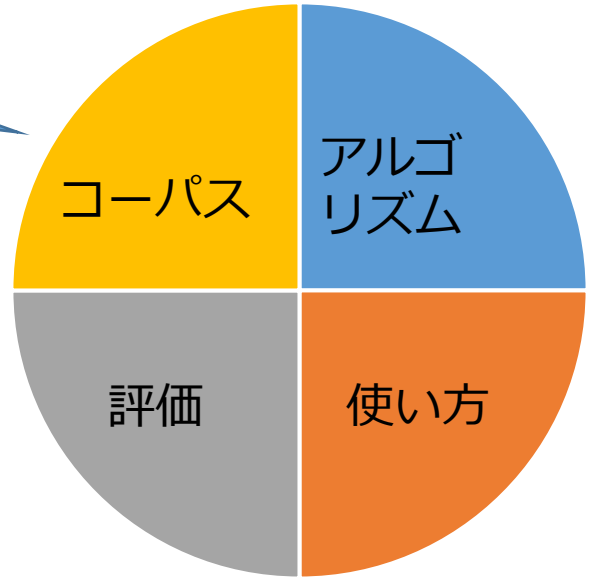
ニューラルネットワークによる高精度アルゴリズム

高精度自動翻訳の社会的普及

NMTでも以前と同じ点

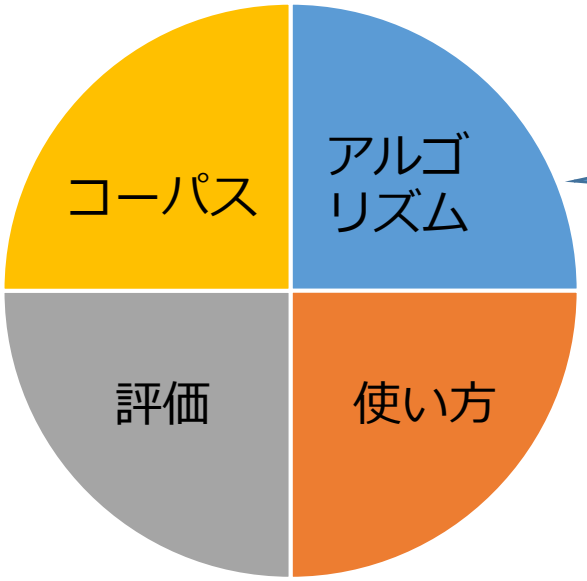
大規模対訳
コーパスが必要

人間による評価
が必要



ＮＭＴアルゴリズムの概要

- 基本訓練
- アダプテーション



ニューラルネットワークによる高精度アルゴリズム

ＮＭＴの訓練とアダプテーション概要

【訓練】 １文ずつＮＭＴモデルのパラメタを調整する

- 大雑把には：
  - 入力文を翻訳
  - 参照訳文と比較
  - 翻訳文と参照訳文の違いに応じてＮＭＴのパラメタを更新
  - 以上を大規模に繰り返す（数億回になることもある）

【アダプテーション】

- 訓練済みＮＭＴモデルに、上記訓練を特定分野データで追加
- 訓練済みモデルをベースにするので、比較的少量データで高精度

NMTの基本構造（第１・２世代ほぼ共通）

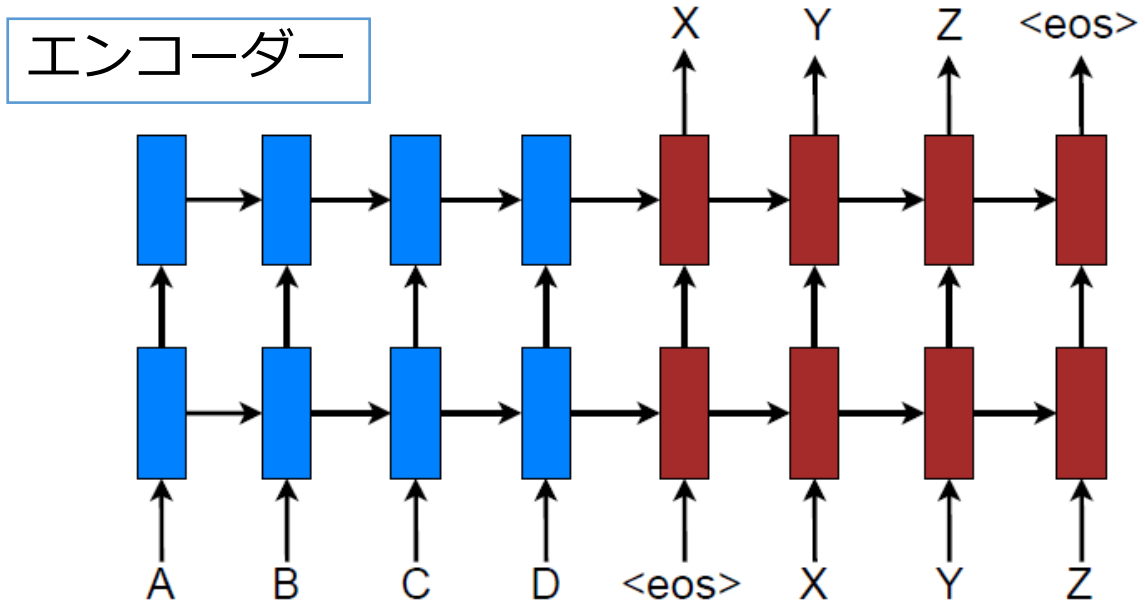


Figure 1: **Neural machine translation** – a stacking recurrent architecture for translating a source sequence A B C D into a target sequence X Y Z. Here, <eos> marks the end of a sentence.

- 系列を入力して、系列を出力
- エンコーダー：入力文の受付
- デコーダー：１単語ずつ出力
- 構造は世代間・モデル間で異なる
  - 第１世代 RNN
  - 第２世代 Transformer

Thang Luong; Hieu Pham; Christopher D. Manning. (2015) *Effective Approaches to Attention-based Neural Machine Translation*. EMNLP

第１世代と第２世代の精度比較

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

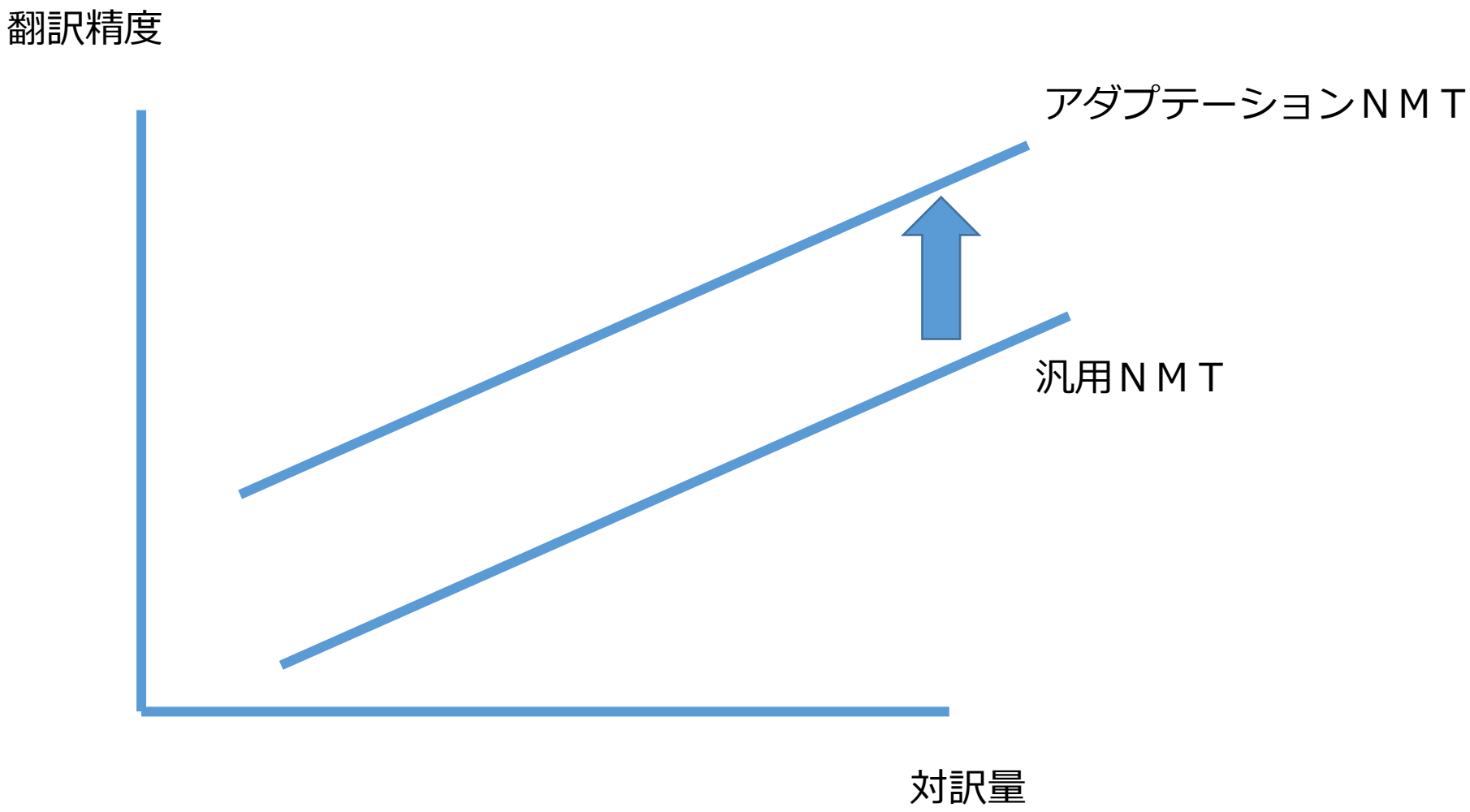
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

翻訳の比較的難しい英独でＢＬＥＵ４ポイント程度の大きな改善

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. (2017) “Attention Is All You Need”. https://arxiv.org/abs/1706.03762

アダプテーション@みんなの自動翻訳

汎用とアダプテーションＮＭＴの関係



製薬NMT：R&D Head Club ＋NICT

- RDHCの8社から提供された320万文対以上の日英対訳データ
- 一昨年度のNICTとアストラゼネカの協業を発展
- 汎用NTを製薬対訳でアダプテーション
- 事業会社によりサービス提供が開始済み

２０１９年１０月７日　https://www.nict.go.jp/info/topics/2019/10/07-1.html

第２世代製薬NMTは高精度（再掲）

	第１世代	第２世代
汎用	２３．８	３０．３
製薬	３９．９	４９．４

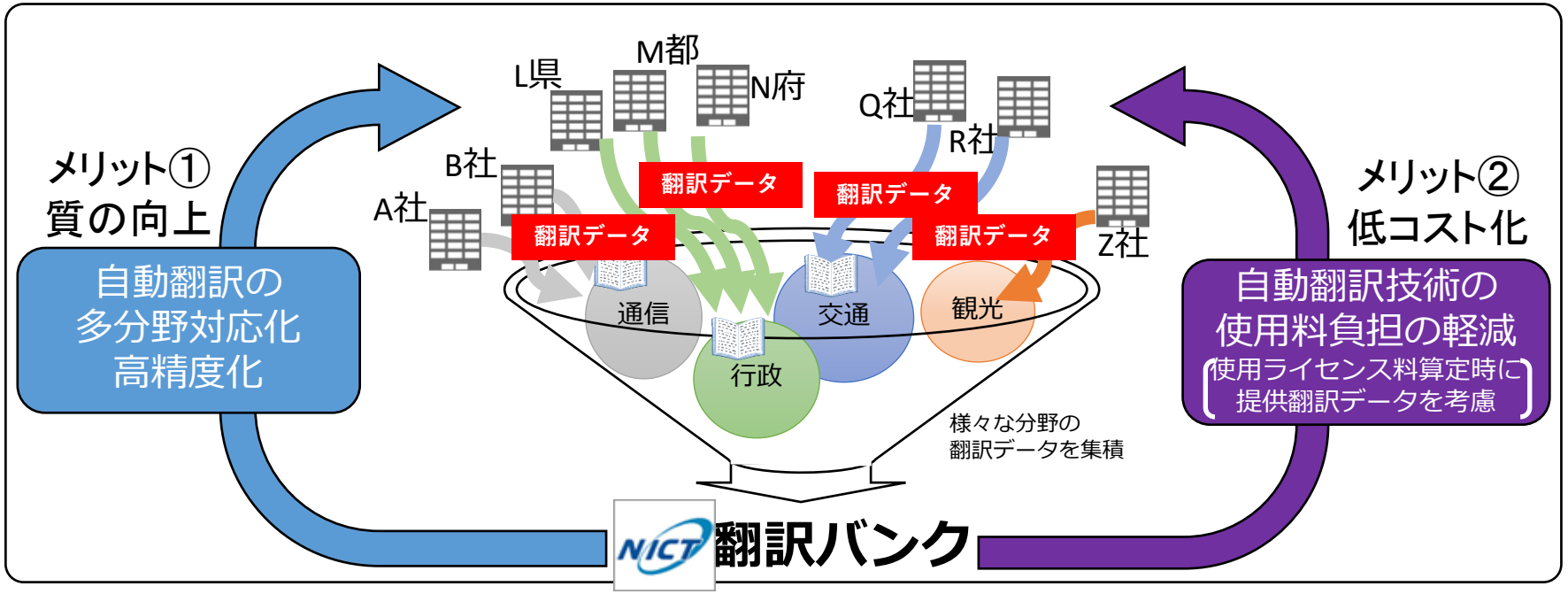
第１世代＜＜第２世代

汎用　＜＜　製薬

自動翻訳尺度BLEUの比較

翻訳バンクで高品質対訳 アダプテーションで高精度NMTを実現

NICTの自動翻訳技術の使用ライセンス料の算定の際に、提供が見込まれる翻訳データを勘案して負担を軽減する仕組みを導入



覚えておいていただきたいこと

- 自動翻訳エンジンは進化します
- アルゴリズムの進化
- 対訳データの進化
- 自動翻訳精度の向上に頭打ち感はありません
- 自動翻訳をみんなで育てましょう！