自動翻訳技術の概要：なにができるか／できるようになってきているか

情報通信研究機構（ＮＩＣＴ）
内山将夫

特許情報シンポジウム

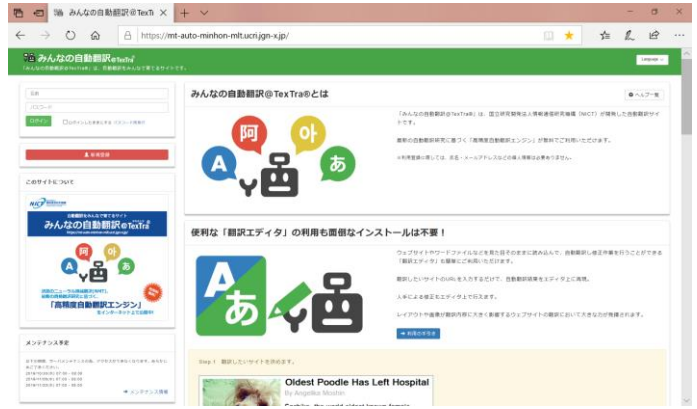
2021年2月26日

本講演の目的

- 自動翻訳技術の概要をお伝えする
- 情報通信研究機構の取り組みの一部をご紹介する

自動翻訳の歴史

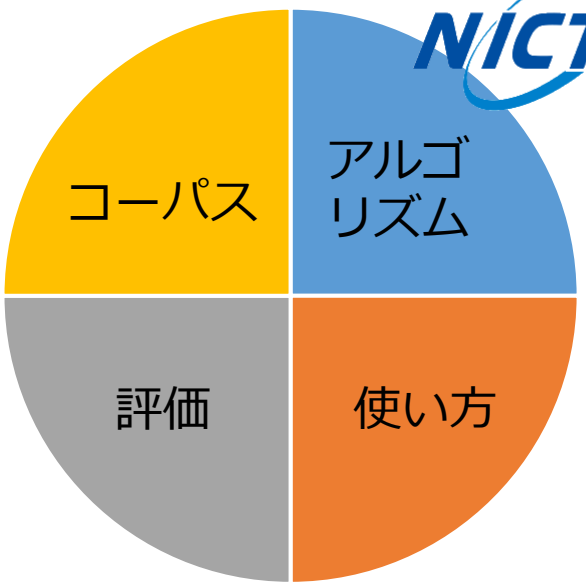
- 計算機が発明されてすぐに自動翻訳の研究が開始(1950年代)
- 自動翻訳の考え自体は 1949 年にWarren Weaverが提案（cf. Wikipedia)
- 半世紀以上の研究開発を経て、自動翻訳が一般に普及
  - みんなの自動翻訳@TexTra等のＷｅｂ翻訳サービス
  - VoiceTra等のスマフォアプリ
  - ポケトーク等の音声翻訳専用機



自動翻訳技術のタイプ

- 規則ベース自動翻訳
  文法規則や辞書を人間が記述
  上記に基づき自動翻訳を実施
- コーパスベース自動翻訳（ＭＴ）
  対訳コーパスから自動翻訳エンジンを自動学習
  任意言語対に対して適用可能
  ニューラル機械翻訳（ＮＭＴ）はこちら

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

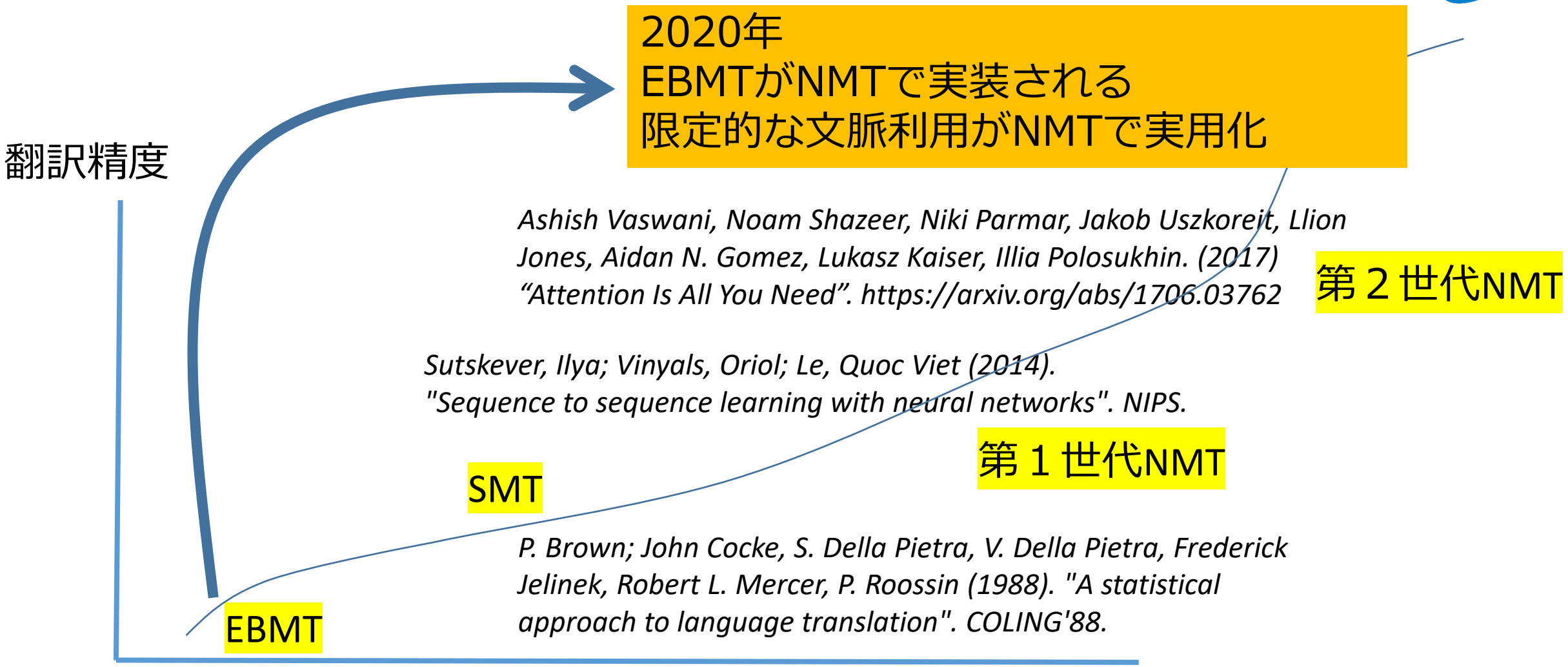
ＭＴの普及により、ＭＴの適切な使い方の周知が必要

対訳コーパスの一例

- Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
- Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
- フランスのパリ、パルク・デ・プランスで行われた2007年ラグビーワールドカップのプールCで、イタリアは31対5でポルトガルを下した。
- អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងប៉ូលCនៃពីធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ាសឌេសប្រីន ក្រុងប៉ារីស បារាំង។
- Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
- ပြင်သစ်နိုင်ငံ ပါရီမြို့ပါဒက်စ် ပရင်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ ဖလား တွင် အီတလီ သည် ပေါတူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားပါသည်။
- Ý đã đánh bại Bồ Đào Nha với tỉ số 31-5 ở Bảng C Giải vô địch Rugby thế giới 2007 tại Parc des Princes, Pari, Pháp.
- อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มC ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
- Natalo ng Italya ang Portugal sa puntos na 31-5 sa Grupong C noong 2007 sa Pandaigdigang laro ng Ragbi sa Parc des Princes, Paris, France.

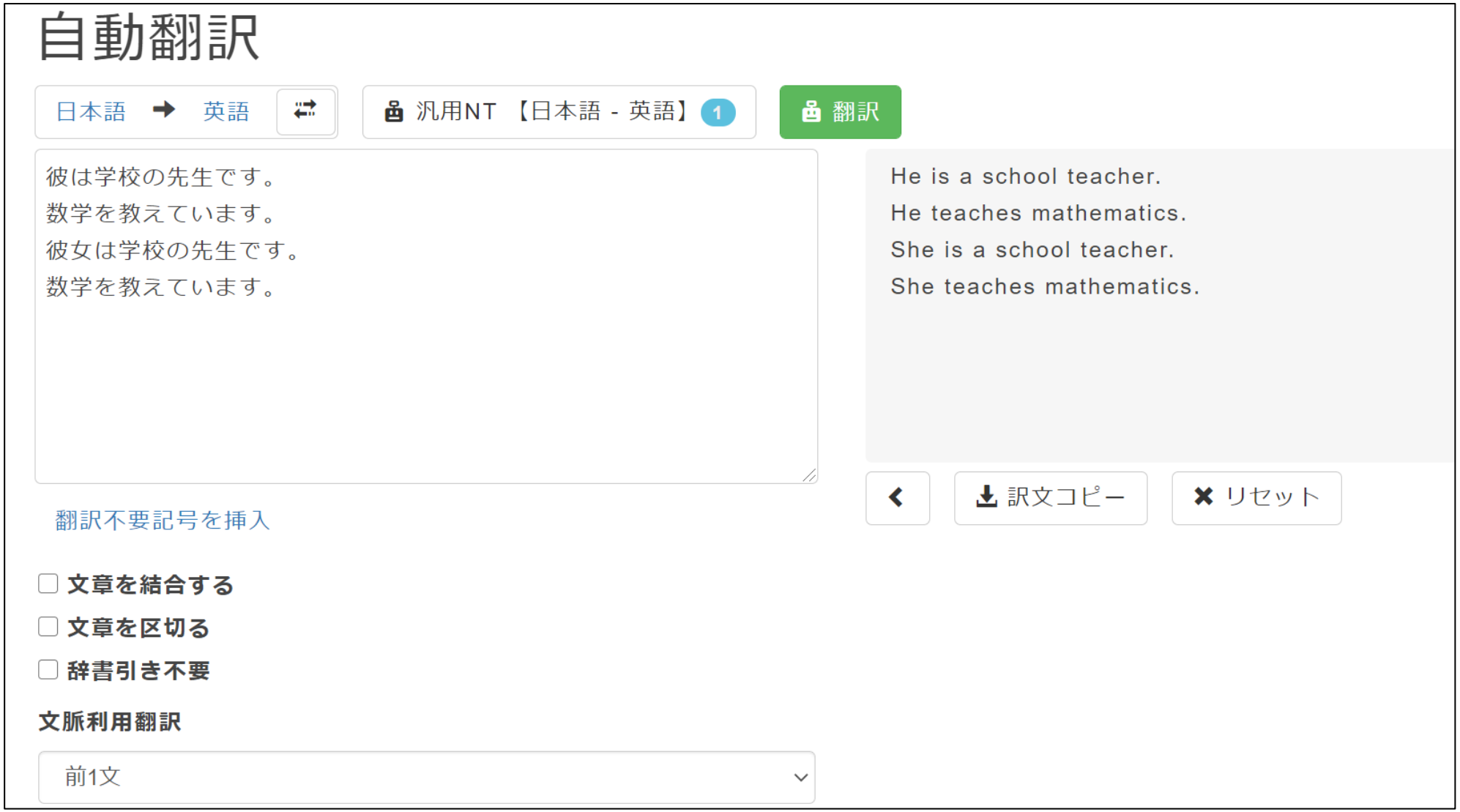
Asian Language Treebank http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

コーパスベースＭＴアルゴリズムの進展



みんなの自動翻訳@TexTra
https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

文脈翻訳の事例


自動翻訳
日本語
英語
汎用NT 【日本語 - 英語】 1
翻訳
彼は学校の先生です。
数学を教えています。
彼女は学校の先生です。
数学を教えています。
He is a school teacher.
He teaches mathematics.
She is a school teacher.
She teaches mathematics.
訳文コピー
リセット
翻訳不要記号を挿入
文章を結合する
文章を区切る
辞書引き不要
文脈利用翻訳
前1文

専門的文章の翻訳例

In Transformer-based neural machine translation (NMT), the positional encoding mechanism helps the self-attention networks to learn the source representation with order dependency, which makes the Transformer-based NMT achieve state-of-the-art results for various translation tasks.
However, Transformer-based NMT only adds representations of positions sequentially to word vectors in the input sentence and does not explicitly consider reordering information in this sentence.
In this paper, we first empirically investigate the relationship between source reordering information and translation performance.
The empirical findings show that the source input with the target order learned from the bilingual parallel dataset can substantially improve translation performance.
Thus, we propose a novel reordering method to explicitly model this reordering information for the Transformer-based NMT.
The empirical results on the WMT14 English-to-German, WAT ASPEC Japanese-to-English, and WMT17 Chinese-to-English translation tasks show the effectiveness of the proposed approach.

Transformer-based Neural Machine Translation(NMT)において,位置符号化機構は自己注目ネットワークが順序依存性を持つソース表現を学習するのを助ける。
これは,Transformer-based NMTが種々の翻訳タスクに対して最先端の結果を達成することを可能にする。しかし,Transformer-based NMTは入力文中の単語ベクトルに連続的に位置の表現を加えるだけで,この文中の情報の順序変更を明示的に考慮しない。
本論文では,まずソース順序変更情報と翻訳性能との関係を経験的に検討する。
経験的発見は,バイリンガル並列データセットから学習したターゲット順序を持つソース入力が翻訳性能を大幅に改善できることを示す。
そこで,Transformer-based NMTに対して,この順序変更情報を明示的にモデル化する新しい順序変更法を提案した。
WMT14English-to-German,WAT ASPEC Japanese-to-English,WMT17Chinese-to-English翻訳タスクの経験的結果は,提案アプローチの有効性を示した。

汎用NT＋でEBMTをお試しください

汎用NT＋は実験用なので技術移転はいたしません。

言語方向	EBMTとして利用されている分野の例
英日	自動車法規、契約、OSS
日英	自動車法規、契約、天気予報、政府関係（外交・運輸安全）
中日	自動車法規
日中	自動車法規

両政府は、前記の了解から又やそれに関連して生ずることがあるいかなる事項についても相互に協議する。

➡

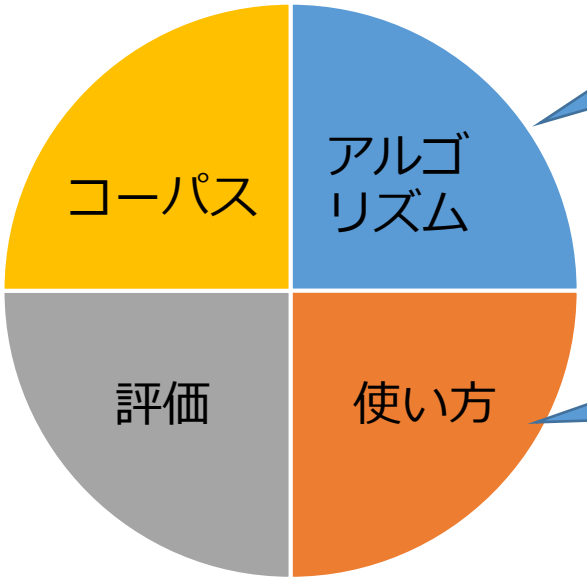
The two Governments will consult with each other in respect of any matter that may arise from or in connection with the foregoing understanding.

对于汽车及其非完整车辆、摩托车和挂车，如果按照 GB 16735 的规定，对已标示的车辆识别代号

➡

自動車及びその不完全車両、オートバイ、トレーラーについて、GB16735の規定に従い、既に表示されている車両識別コードを新たに表示する又は変更する場合、「はい」と記入するものとする。

NMTにより新しくなった点



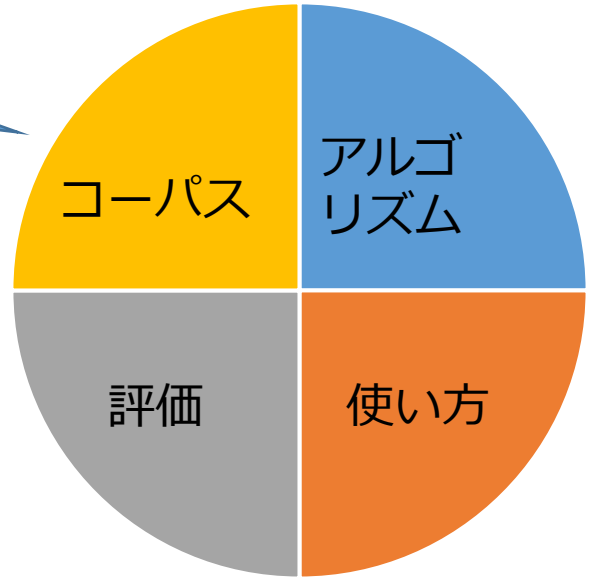
ニューラルネットワークによる高精度アルゴリズム

高精度自動翻訳の社会的普及

NMTでも以前と同じ点

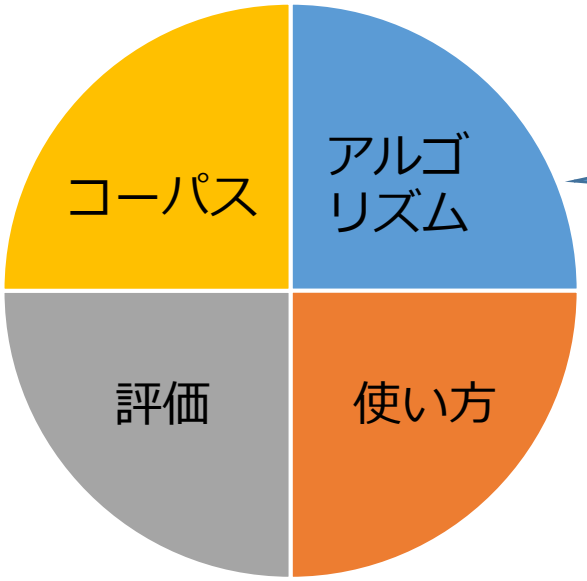
大規模対訳
コーパスが必要

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
- EBMT



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）

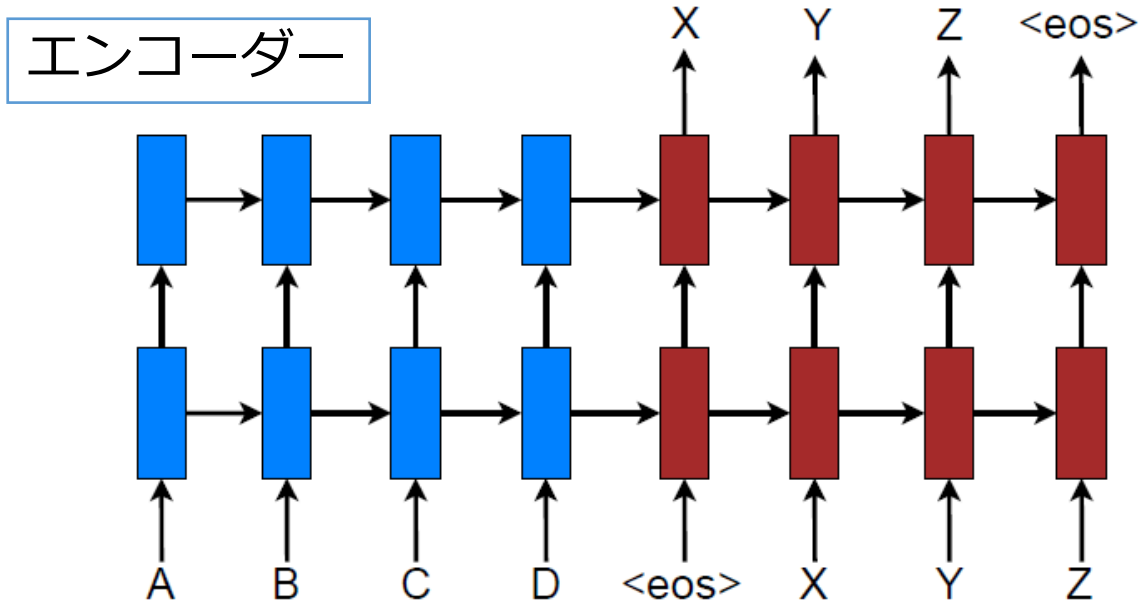


Figure 1: **Neural machine translation** – a stacking recurrent architecture for translating a source sequence A B C D into a target sequence X Y Z. Here, <eos> marks the end of a sentence.

- 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世代の自動評価尺度BLEUの比較

	汎用日英	汎用英日	特許日英	特許英日	
2017	22.1	24.2	40.7	41.7	第1世代
2018	21.2	24.5	40.0	41.6	第1世代
2019	20.8	23.2	37.7	40.1	第1世代
2020	27.6	31.5	45.6	46.8	第2世代
2021	28.1	31.0	47.3	46.5	第2世代

汎用・特許ともに、
第1世代(RNN)から第2世代(Transformer)にかけて、
大きくBLEU値が向上

訓練・アダプテーション・EBMT

【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】（fine tuning とも言います）

- 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

【EBMT】（NICT開発・詳細未発表）

- 入力文と類似した対訳文をデータベースから検索
- 十分に類似した文があるときには、それを参考に自動翻訳
- 類似文がない場合には、ベースのNMTで自動翻訳

アダプテーションとEBMTの特徴

◉ **アダプテーション＋EBMT**

ベースモデルをアダプテーションしたモデルを利用してEBMTを実施します。訓練データが 1 万文以上のときにお勧めします。

○ **アダプテーション**

ベースモデルを訓練データでアダプテーションします。訓練データと似ている文の自動翻訳精度向上が見込めますが、訓練データと似ていない文の自動翻訳精度がベースモデルより低下する場合があります。訓練データが 1 万文以上で、少しノイズがあるときにお勧めします。当初から「アダプテーション」として当サイトで提供していた訓練方法です。

○ **EBMT**

Example-based Machine Translationです。訓練データ中に入力文と似ている文がある場合には、それを参考に自動翻訳します。入力文との類似文がない場合はベースモデルで翻訳します。訓練データが 1 万文未満のときにお勧めします。

みんなの自動翻訳からお試しできます

アダプテーションの有効性の事例

自動車法規文の自動翻訳をニューラル技術で高精度化
～トヨタとの共同研究を通じ、英日・中日翻訳の実用度が向上～

- 自動車業界からトヨタが翻訳バンクに協力、翻訳データを提供
- 自動車法規を対象とした翻訳をニューラル英日翻訳システムで24%実用度向上

法規原文	人による翻訳	ニューラル翻訳
This Australian Design Rule prescribes requirements for the number and mode of installation of lighting and light signalling devices on motor vehicle other than L-group vehicles.	本オーストラリア設計規則はLグループ車両以外の自動車への灯火および灯火信号装置の数と取り付け方法に関する要件を規定する。	本オーストラリア設計規則は、Lグループ車両以外の自動車への灯火および灯火信号装置の数および取り付け方法に関する要件を規定する。

２０１９年４月２３日　https://www.nict.go.jp/press/2019/04/23-1.html

アダプテーションで専門用語に適応

原文	第１世代汎用ＮＭＴ	第２世代汎用ＮＭＴ	第１世代金融ＮＭＴ	第２世代金融ＮＭＴ
本取引は、相互売買による取引実績が豊富な本投資法人ならではの取引であると考えており、本投資法人は今後も戦略的かつ継続的に資産入替を実施していく方針です。	This transaction is considered to be a transaction unique to this investment corporation, which has a rich record of transactions through mutual trade. This investment corporation will continue to change assets in a strategic and ongoing manner.	We believe that this transaction is unique to the Investment Corporation, which has a rich track record of transactions through mutual transactions. The Investment Corporation will continue to implement this transaction strategically and continuously.	The Transaction is considered as a transaction that is unique to the Investment Corporation, which has a wealth of transaction experience through mutual transaction. The Investment Corporation will continue to implement strategic and continuous asset replacement in the future.	The Investment Corporation believes that the Transaction is unique to the Investment Corporation, which has a wealth of transaction records through mutual transactions. Therefore, the Investment Corporation will continue to implement strategic and continuous asset replacement.

AAMT 2019, Tokyo: パネルディスカッション「機械翻訳もある総合的な翻訳サービスの模索～金融・IR分野を例に～」の内山のスライドの１枚

金融分野でのアダプテーション・EBMT

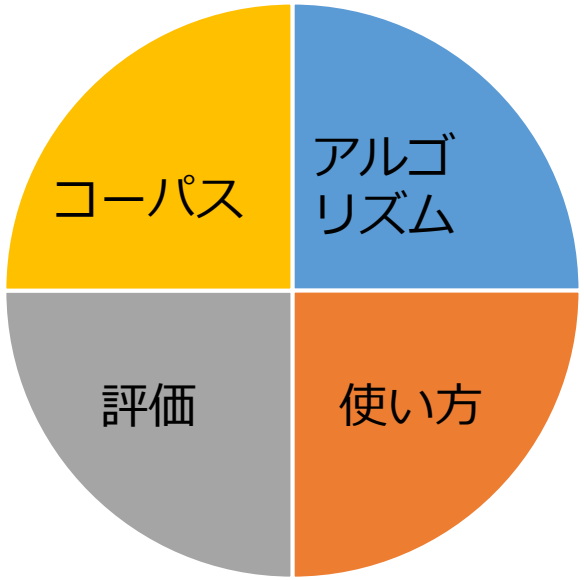
	汎用	ADAPT	EBMT	ADAPT+EBMT
訓練と重複あり1000	31.0	38.1	54.5	55.2
訓練と重複無し514	29.3	35.2	47.7	49.1
訓練と重複のみ416文	34.4	44.1	68.3	67.5

適時開示文書は、異なる文書であっても重複する文が多いので、EBMTの効果が著しく高い

ＭＴの評価の概要

- ＭＴの評価はＭＴを改善する原動力
- 誤訳報告⇒誤訳の改善
- 適切な評価⇒適切なＭＴ

人間による評価が必要



自動評価

- 簡便
- ＭＴ訳と参照訳を比較
- 評価値を自動計算
- 開発にはとても役に立つ！
  - システム改修後、自動評価値に大きな変化がないことを確認
  - 大きな数値差（ＢＬＥＵなら５程度）があれば品質差もほぼある

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」

自動評価尺度BLEUでの比較

	第１世代	第２世代
汎用	２５．１	２９．０
金融	３４．６	３８．４

第１世代＜＜第２世代

汎用　＜＜　金融

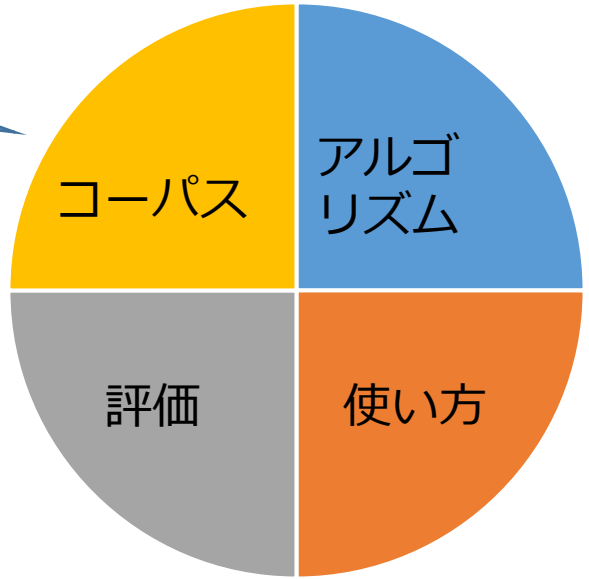
AAMT 2019, Tokyo: パネルディスカッション「機械翻訳もある総合的な翻訳サービスの模索～金融・IR分野を例に～」の内山のスライドの１枚より引用
日英翻訳の例について、みんなの自動翻訳でＢＬＥＵを計算

人間評価

- 原文とＭＴ訳を人間が比較
- 人間が誤訳かを分析
- ＭＴの研究開発に必須
- 第２世代ＮＭＴでは、おおむね正しい翻訳ができるようになり、以前は見逃されていた誤訳が目立ち始めた
  - 「ＮＩＣＴ」⇒「ＮＩ ＣＴ」（なぜか空白が挿入）
- ピンポイントな誤訳を検出する自動評価方法は知られていない
- ピンポイントな誤訳は、ピンポイントなルールなどで改善

NMTでの対訳コーパスの重要性

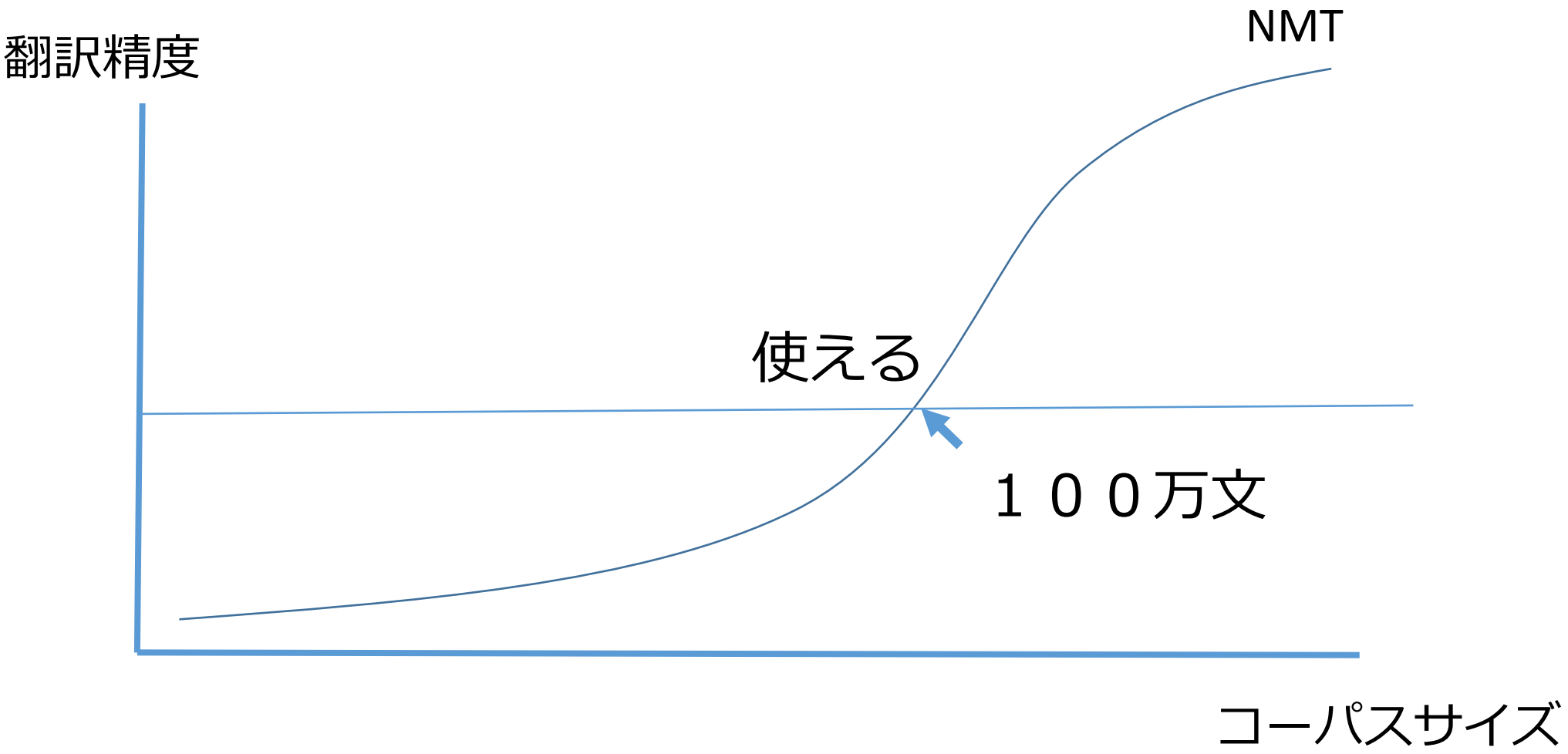
大規模対訳
コーパスが必要



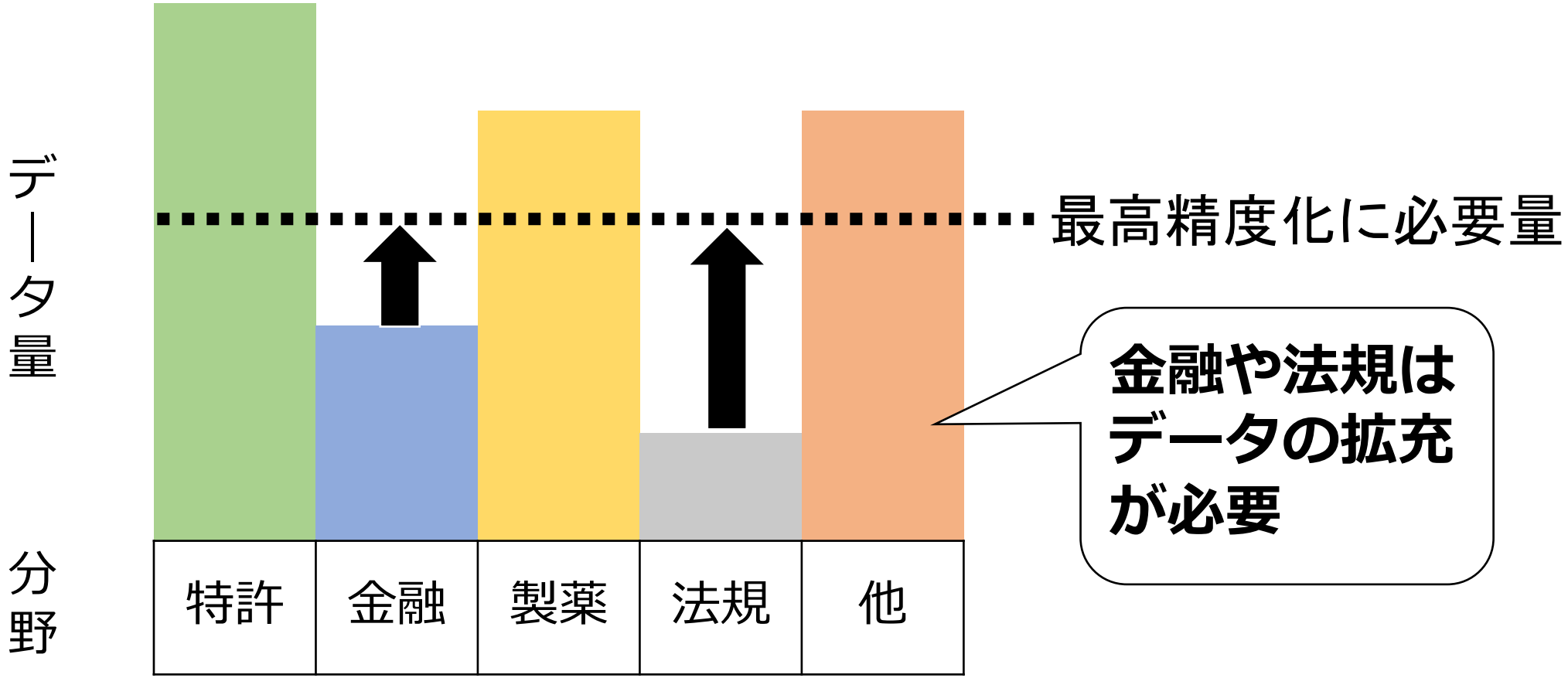
対訳コーパスの一例（再掲）

- Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
- Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
- フランスのパリ、パルク・デ・プランスで行われた2007年ラグビーワールドカップのプールCで、イタリアは31対5でポルトガルを下した。
- អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងប៉ូលCនៃពីធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ាសឌេសប្រិន ក្រុងប៉ារីស បារាំង។
- Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
- ပြင်သစ်နိုင်ငံ ပါရီမြို့ ပါဒက်စ် ပရင့်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ ဖလား တွင် အီတလီ သည် ပေါ်တူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားပါသည် ။
- Ý đã đánh bại Bồ Đào Nha với tỉ số 31-5 ở Bảng C Giải vô địch Rugby thế giới 2007 tại Parc des Princes, Pari, Pháp.
- อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มC ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
- Natalo ng Italya ang Portugal sa puntos na 31-5 sa Grupong C noong 2007 sa Pandaigdigang laro ng Ragbi sa Parc des Princes, Paris, France.

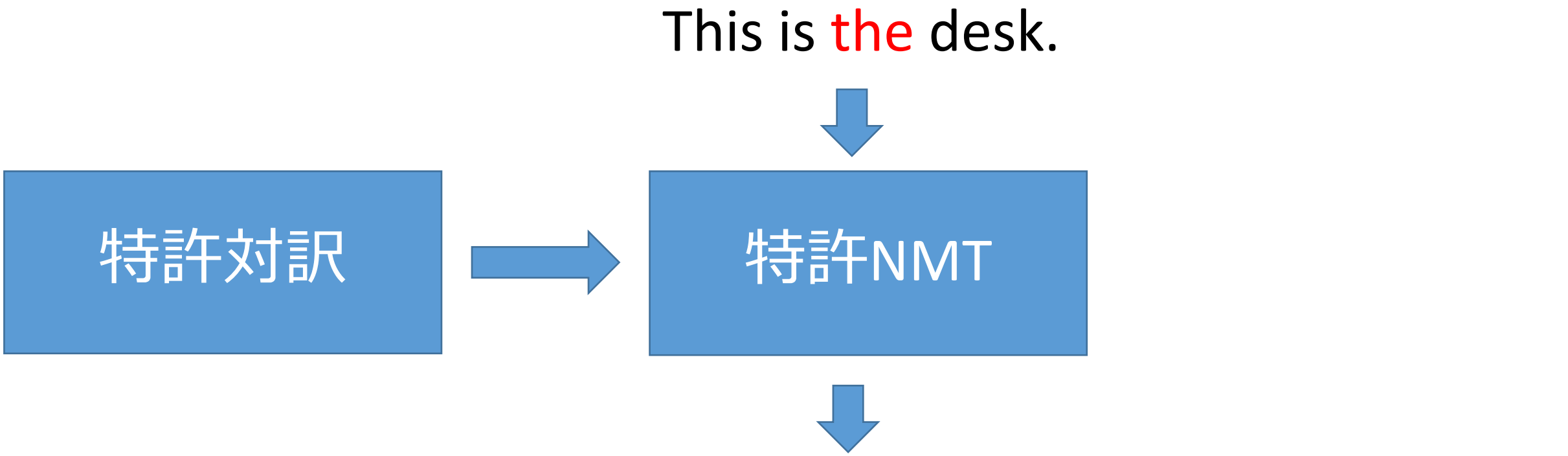
大規模対訳コーパスでＮＭＴの性能向上



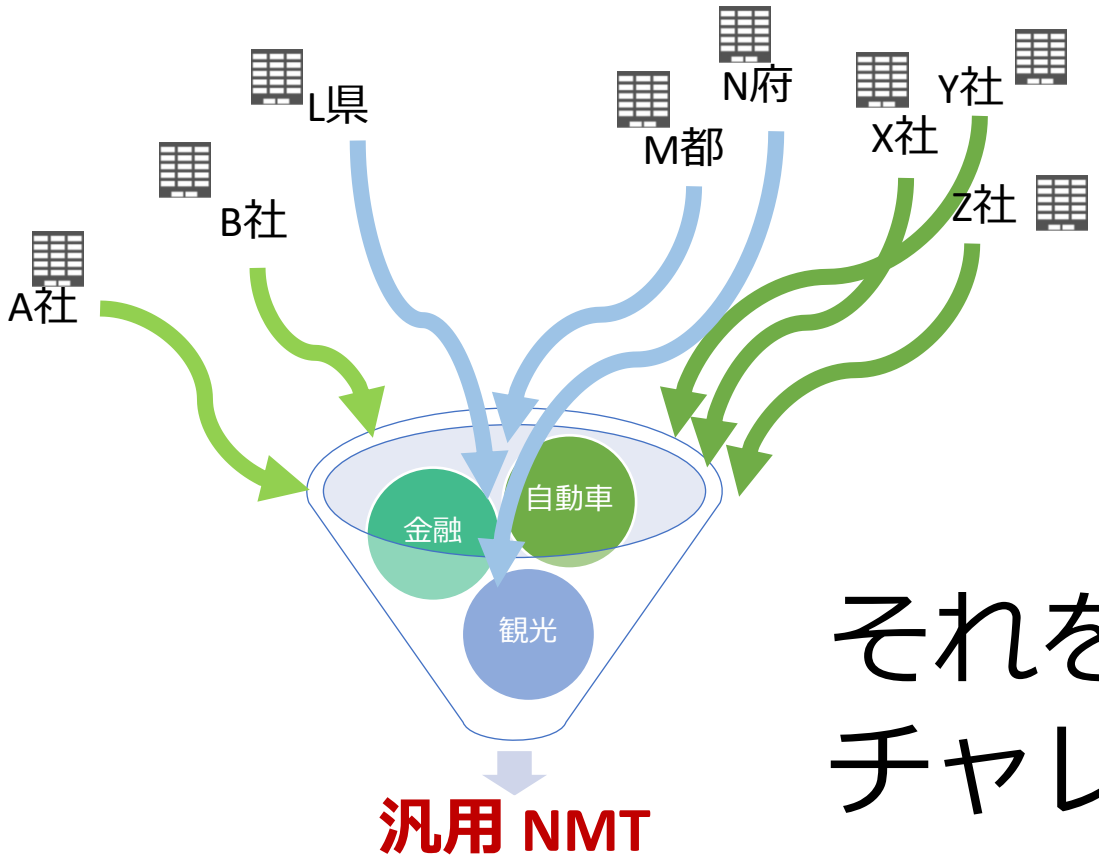
高精度ＮＭＴには対訳データが重要



分野にマッチした対訳が重要
特許対訳は普通の文には不向き



汎用対訳コーパスが汎用ＮＭＴには必要

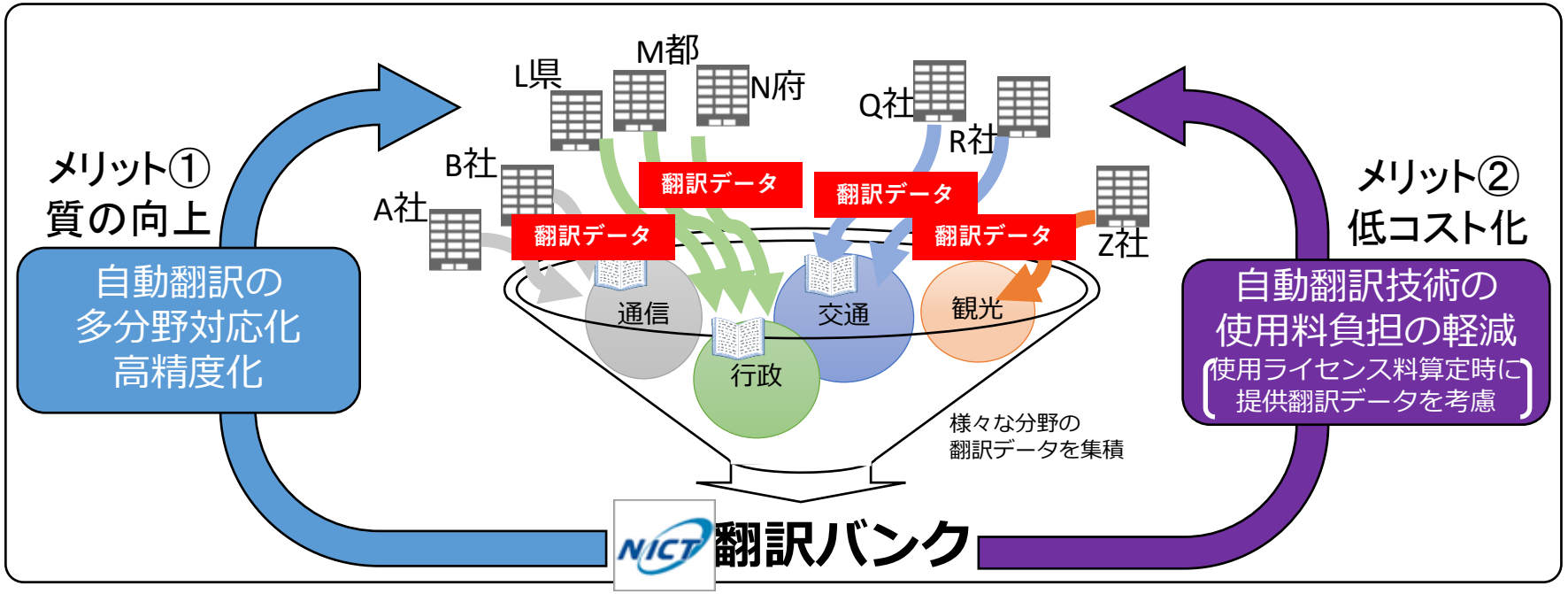


対訳コーパスは
散らばっている

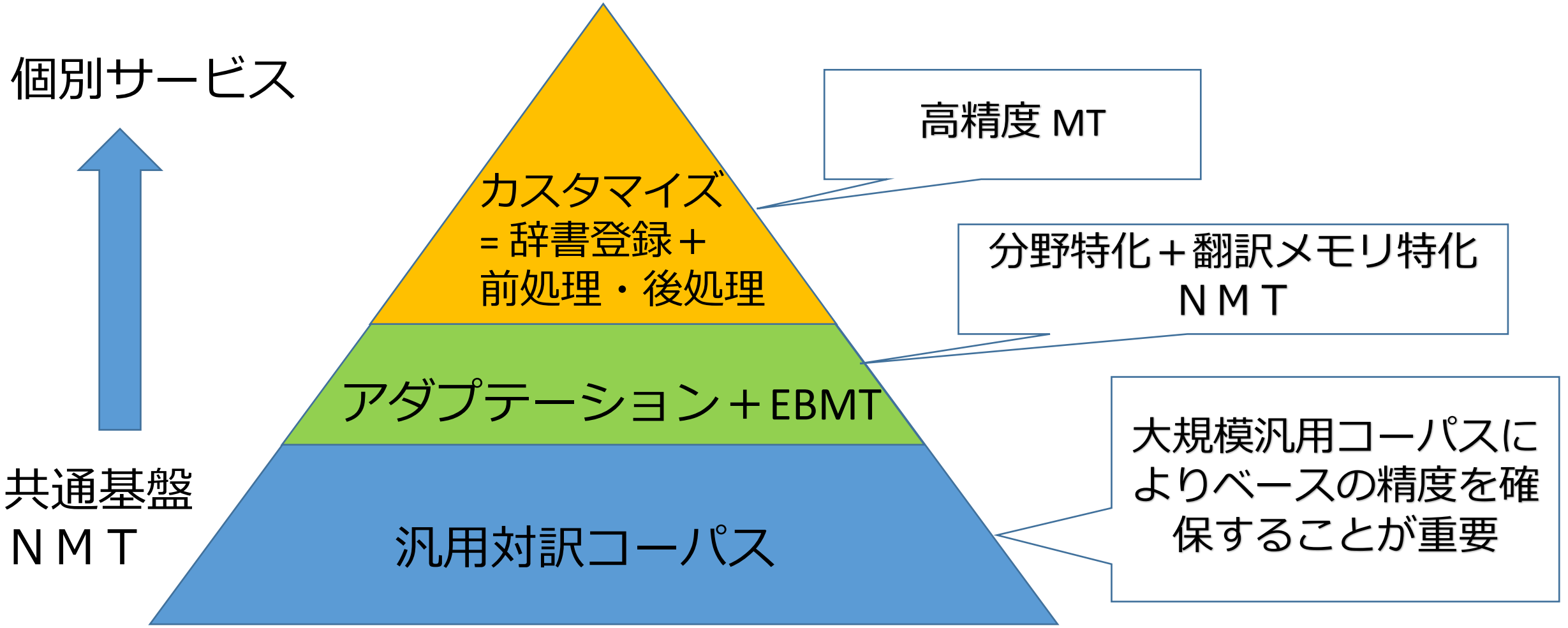
それを集めるのが
チャレンジ

翻訳バンクで高品質対訳データ

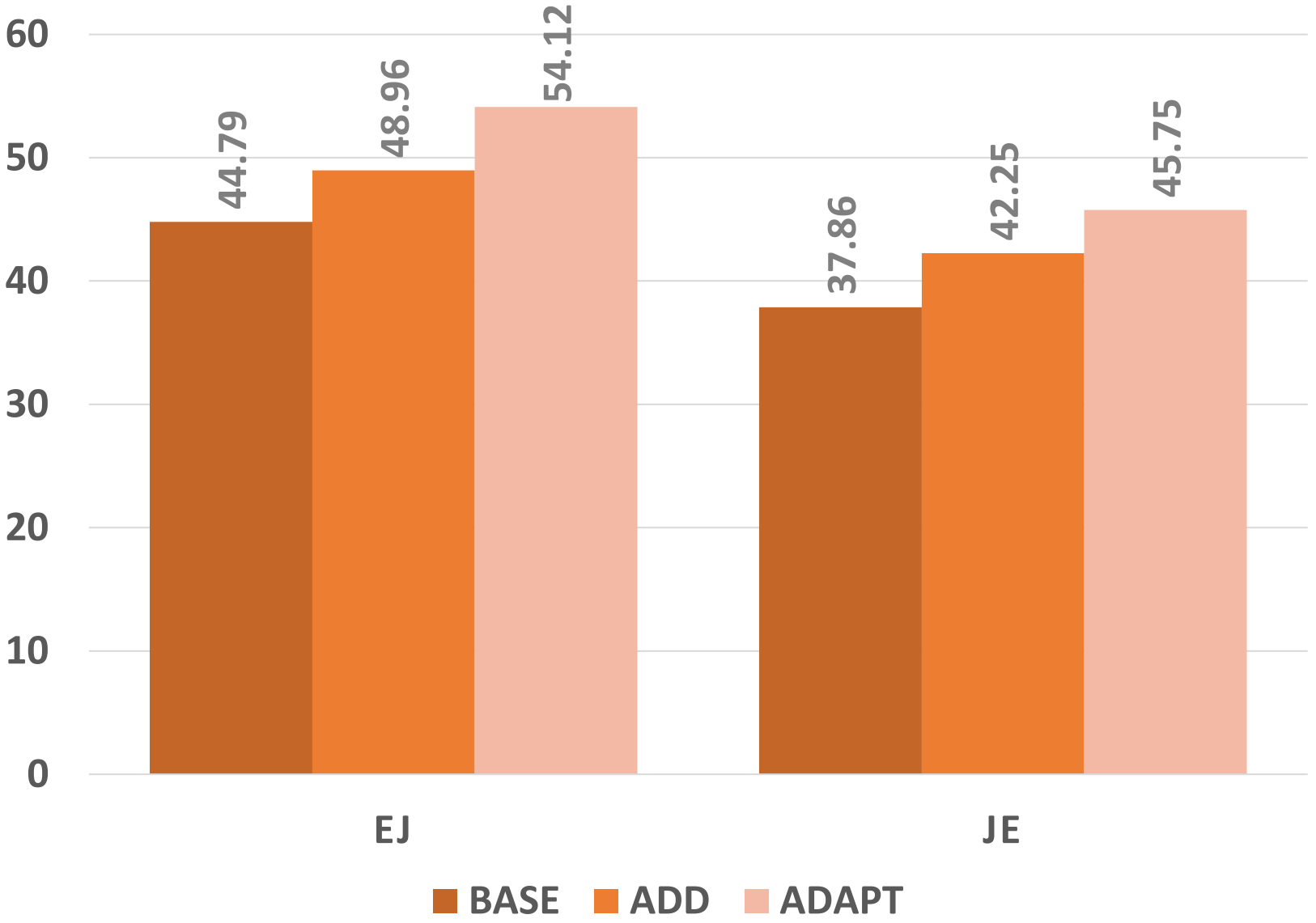
NICTの自動翻訳技術の使用ライセンス料の算定の際に、提供が見込まれる翻訳データを勘案して負担を軽減する仕組みを導入



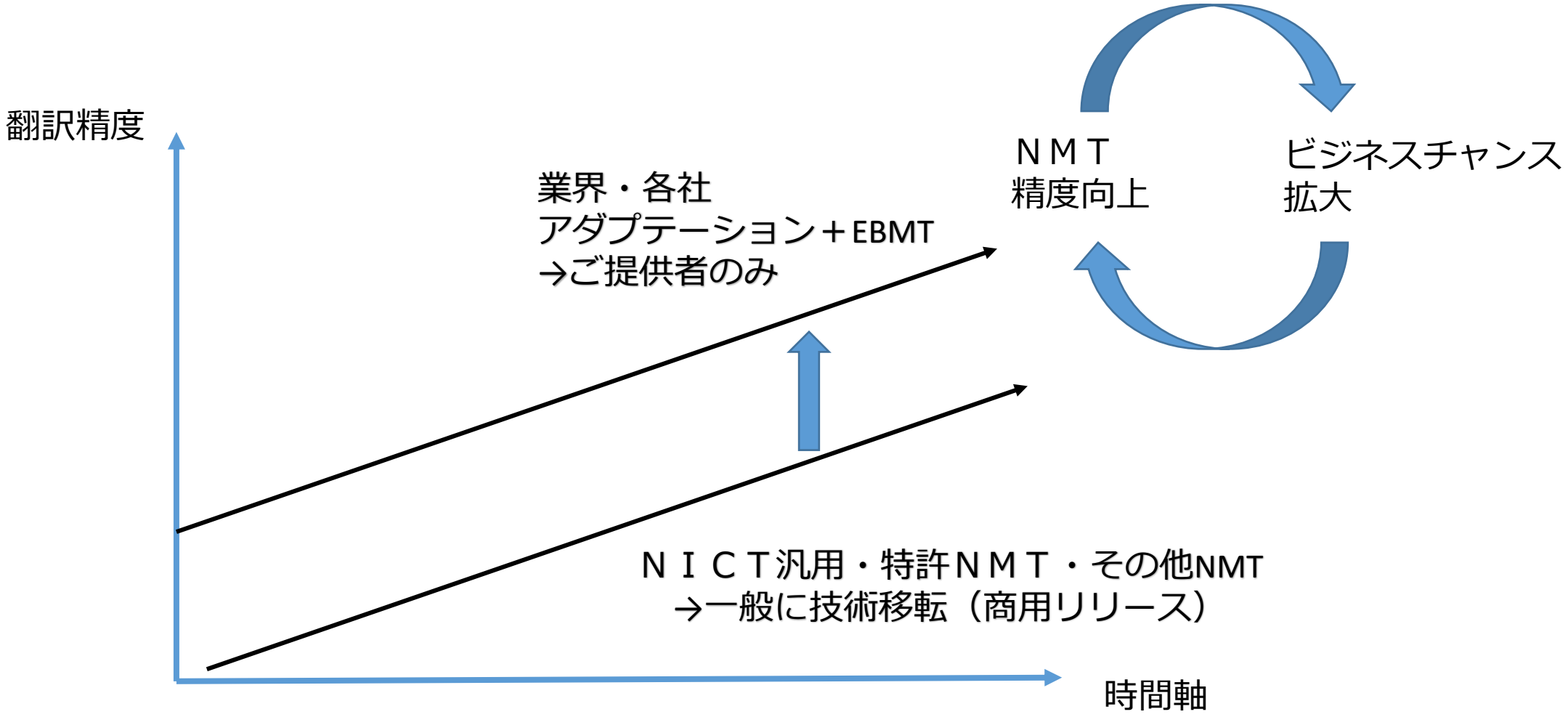
高精度MT = 汎用NMT + アダプテーション +EBMT+カスタマイズ



翻訳精度BLEUの2段階向上



対訳は２度使う


翻訳精度
業界・各社
アダプテーション＋EBMT
→ご提供者のみ
ＮＭＴ
精度向上
ビジネスチャンス
拡大
ＮＩＣＴ汎用・特許ＮＭＴ・その他NMT
→一般に技術移転（商用リリース）
時間軸

翻訳バンクへのご協力方法

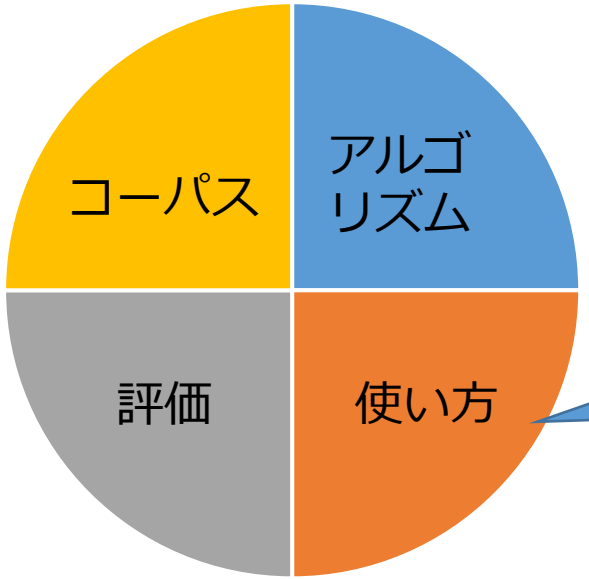
1 『みんなの自動翻訳@TexTra®』の利用

2 NICTとの翻訳データの提供に関する契約の締結

3 ライセンス契約に伴う翻訳データの提供

翻訳データ量でライセンス料の負担軽減

NMTの安全な使い方



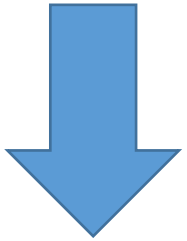
高精度自動翻訳の社会的普及

実用化の一方、誤訳の影響の拡大可能性

自動翻訳の特性を理解した
適切な使い方が必要

ＭＴの安全な使い方

前提：自動翻訳においては、自動翻訳結果の訳文の正確性は保証されません。



どのようにＭＴを使うかが重要

＊ＭＴサービスを利用するときには、その利用規約を参照してください。

自分がＭＴを使う場合

- 自動翻訳結果に間違いが含まれる可能性を理解する
- 理解したうえで、利用する
- 英語の特許等を日本語で読む
- 日本語のメールを英語に自動翻訳して編集する
- 両言語がわかる人がＭＴを使えば、非常に便利で安全な道具

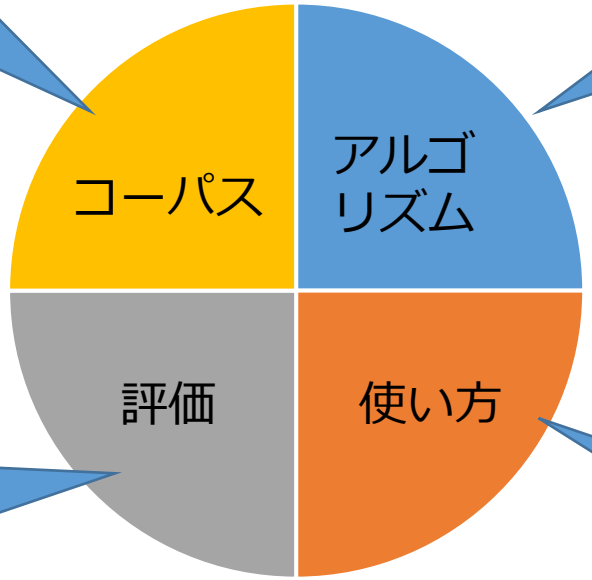
ＭＴの出力を他人が使う場合

- ＭＴの出力を誰か（企業・ボランティア等）が修正する場合
  - 修正する企業等を信頼して利用する
- ＭＴの出力をそのまま利用する場合
  - 何とかして翻訳を正確にしたい
  - 主語・述語が明瞭な文章を入力する。
  - 翻訳結果を再度日本語に翻訳して、意味が一致するかを確かめる
  - 複数の自動翻訳エンジンの翻訳結果が一致するかを確かめる
  - 日本語入力を言い換えた日本語入力について上記を実施する
  - 入力文とＭＴ訳文の双方を利用者に提示する。
  - ＭＴの出力のため、間違いが含まれる可能性があることを知らせる。

まとめ

翻訳バンクにご協力
をお願いします！

第２世代ＮＭＴ
＋アダプテーション
＋EBMT



適切な評価
⇒適切なＭＴ

便利だが
使い方に注意が必要