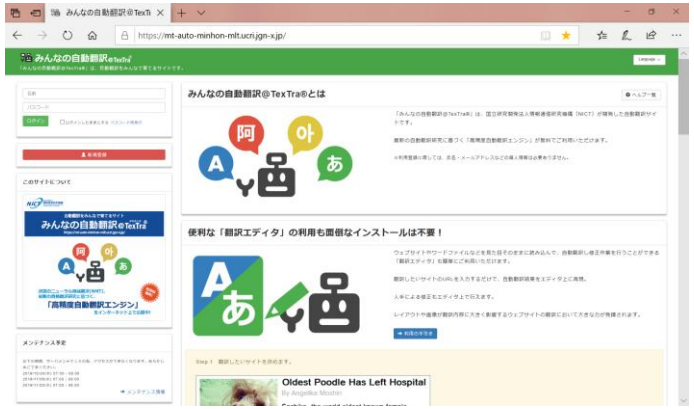
翻訳バンクと アダプテーションによる ＮＭＴ超高精度化

内山将夫
情報通信研究機構（ＮＩＣＴ）
2019 特許・情報フェア＆コンファレンス

自動翻訳の歴史

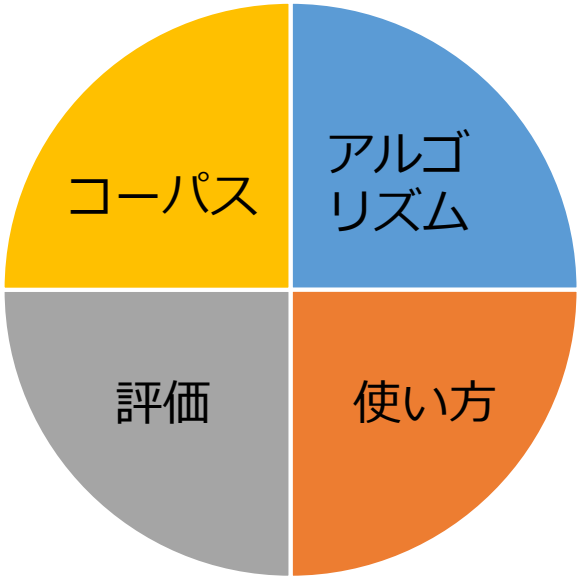
- 計算機が発明されてすぐに自動翻訳の研究が開始(1950年代)
- 自動翻訳の考え自体は 1949 年にWarren Weaverが提案（cf. Wikipedia)
- 半世紀以上の研究開発を経て、自動翻訳が一般に普及
  - みんなの自動翻訳@TexTra等のＷｅｂ翻訳サービス
  - VoiceTra等のスマフォアプリ
  - ポケトーク等の音声翻訳専用機



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

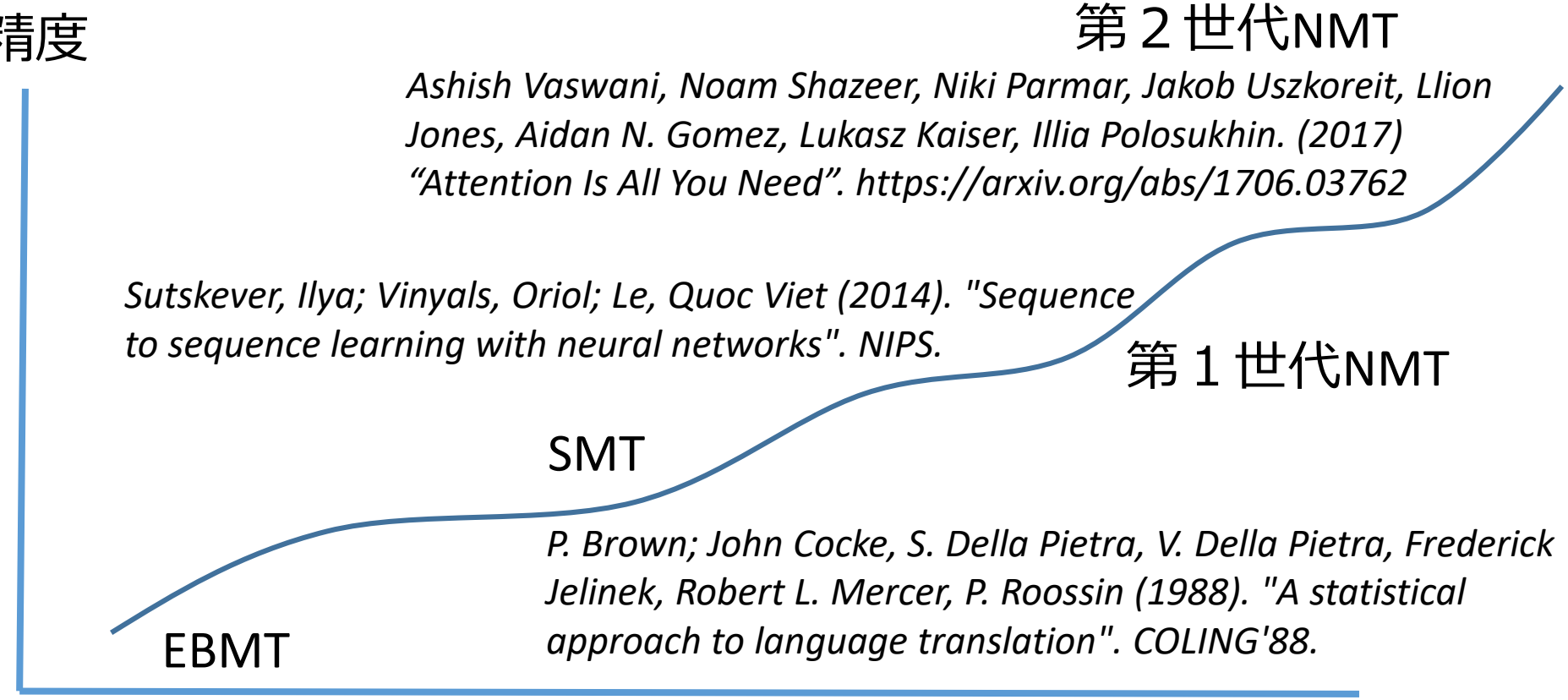
ＭＴの普及により、ＭＴの適切な使い方の周知が必要

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 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.

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



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/

２０１９年はＮＭＴが更に高精度に！

	トピック	リリース主体
３月	オール・ジャパンの枠組みを活用「みんなの自動翻訳@ＫＩ（カスタム版）」の提供を開始	川村インターナショナル
４月	機械翻訳サービスの和文英訳がプロ翻訳者レベルに、 英文和訳はTOEIC960点レベルを達成	みらい翻訳
４月	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～	NICT
５月	特許庁から受注した機械翻訳システムの稼動開始について ～NICTが開発した最新のニューラル機械翻訳エンジンの採用により、正確で自然な翻訳を実現 ～	東芝デジタルソリューションズ
７月	AI翻訳サービス「T-tact AN-ZIN」を発表 「みんなの自動翻訳@TexTra®」が商用利用可	十印
９月	IR・金融分野向け自動翻訳エンジンの性能向上を確認 ～日本財務翻訳との共同研究により実用化へ～	NICT
１０月	大規模翻訳データによる製薬業界向けAI自動翻訳システムの最適化 ～情報通信研究機構（NICT）とR&D Head Clubの共同開発～	NICT

ＮＭＴ高精度化の必要条件

①アルゴリズム
第２世代アルゴリズムTransformerが実用化
汎用NT&特許NT@みんなの自動翻訳

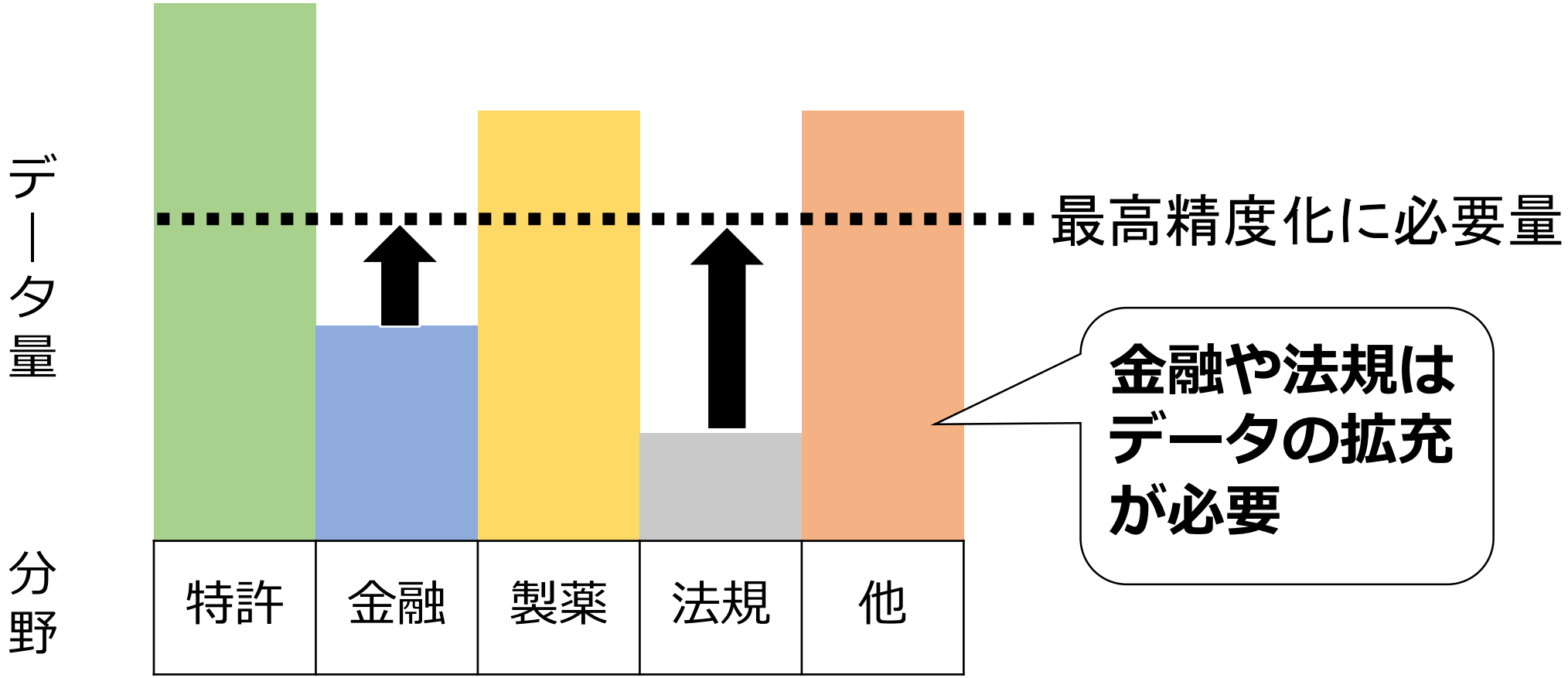
②対訳データの量
汎用・特許ＮＭＴの訓練には数千万文以上の対訳文が必要

③対訳データの質
高品質対訳データ⇒高品質ＮＭＴ
低品質対訳データ⇒低品質ＮＭＴ

①アルゴリズム
第２世代のほうが原文に忠実

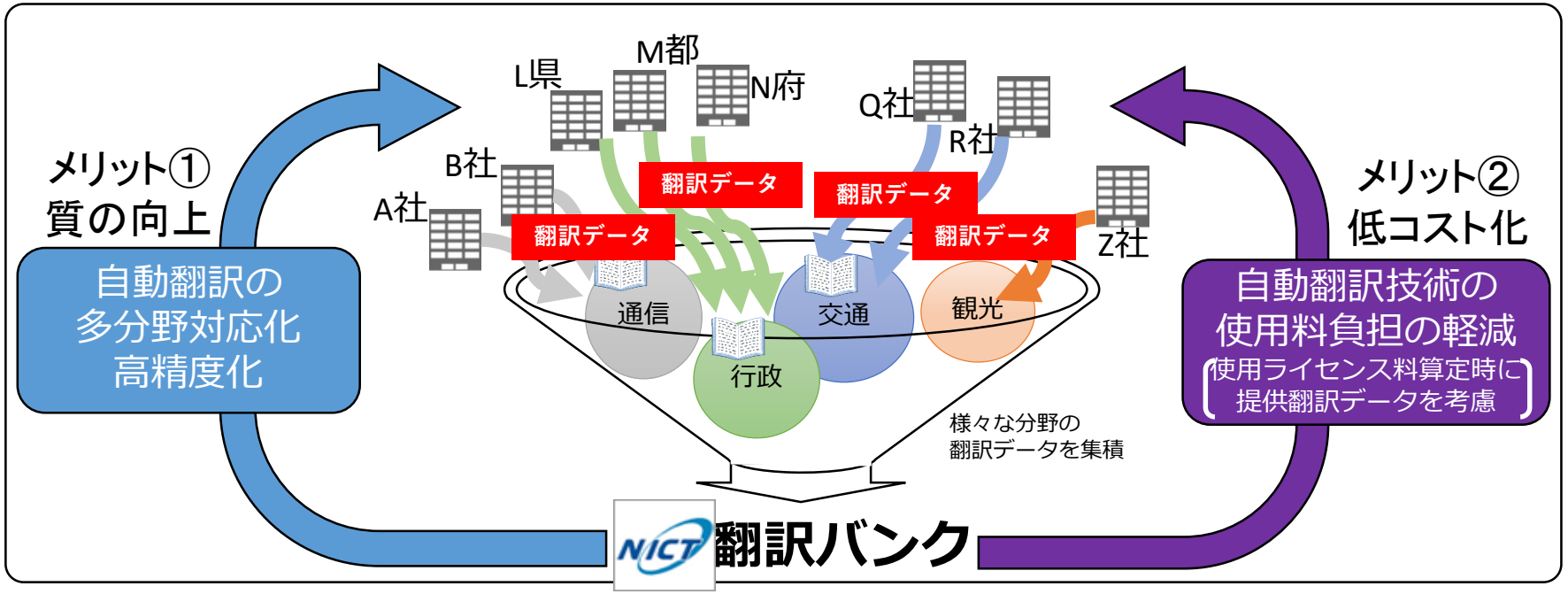
- すなわち本発明の電池は、正電極、負電極およびセパレーターを基本構成要素とする電池において、前記セパレーターが合成樹脂微細多孔膜からなり、該微細多孔膜が線量が０．１～１０Ｍｒａｄの電離放射線処理されていることを特徴とする。
- 【第２世代】 That is, the cell of the present invention comprises a positive electrode, a negative electrode and a separator as basic components, wherein the separator is made of a synthetic resin microporous film, and the microporous film is treated with ionizing radiation at a dose of 0.1 ～ 10 Mrad.
- 【第１世代】That is, the battery of the present invention is characterized in that the separator is made of a synthetic resin microporous membrane in a battery comprising a positive electrode, a negative electrode and a separator as basic components, and the fine porous membrane is subjected to ionizing radiation treatment with a dose of 0.1 ～ 10 Mrad.

②高精度ＮＭＴには対訳データが重要



③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



ＮＭＴの訓練とアダプテーション

- ＮＩＣＴ汎用ＮＭＴ（ベースＮＭＴ）

大量の対訳文について次の訓練を繰り返し実施：

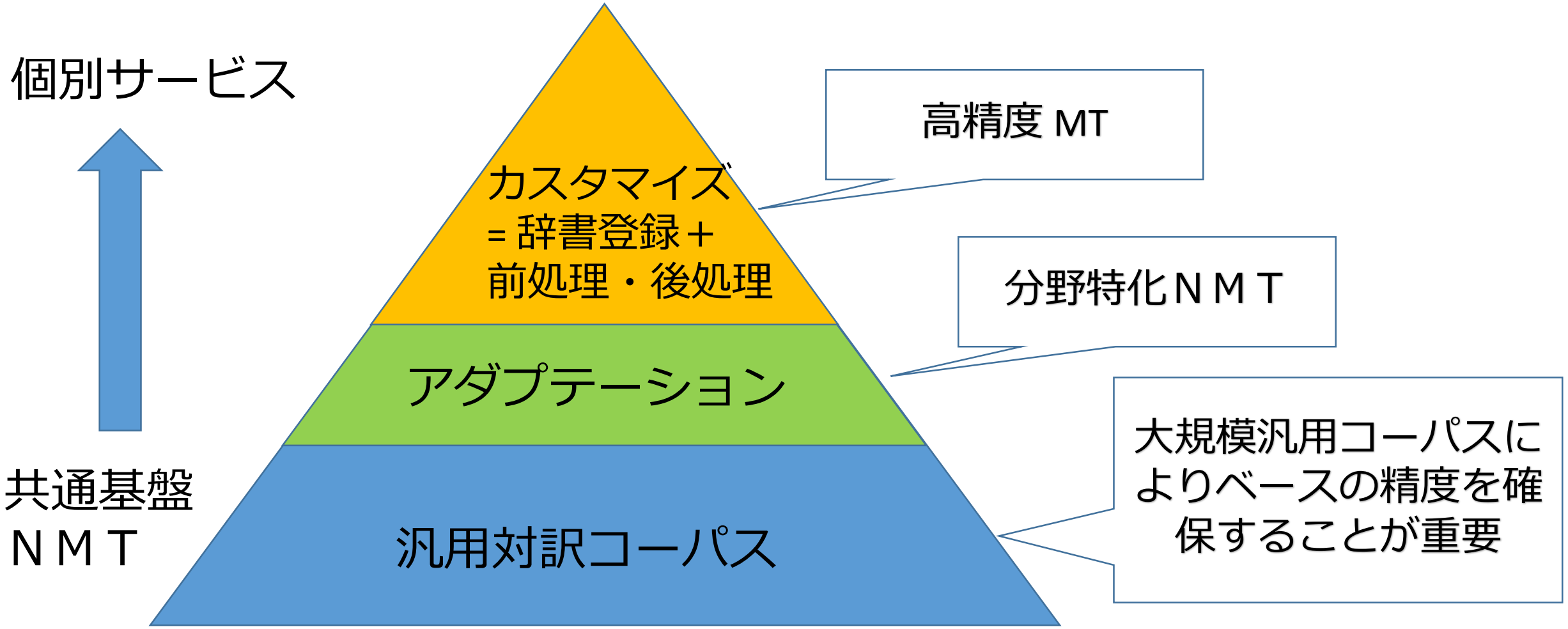
現時点のNMTで原文を翻訳してみて、その結果が参照訳文と異なる度合いに応じて、NMTのパラメタを調整する

- アダプテーションＮＭＴ

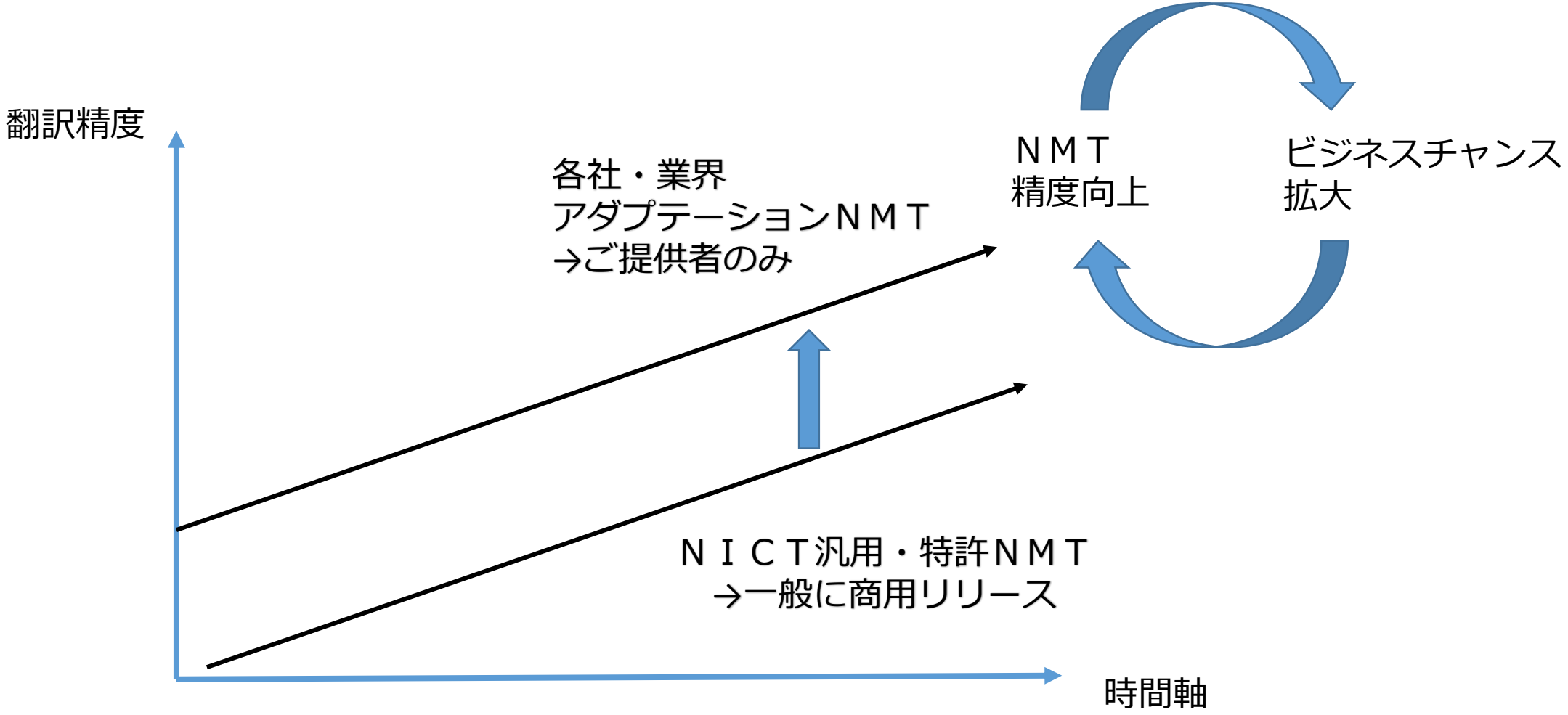
ＮＩＣＴ汎用ＮＭＴをベースに、翻訳対象分野の対訳データについて追加パラメタ調整を実施

⇒比較的少量対訳データで更なる高精度化

高精度MT = 汎用NMT + アダプテーション +カスタマイズ



対訳は２度使う


翻訳精度
各社・業界
アダプテーションＮＭＴ
→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
→一般に商用リリース
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

みんなの自動翻訳デモ


みんなの自動翻訳@TexTra®
「みんなの自動翻訳@TexTra®」は、自動翻訳をみんなで育てるサイトです。
Language
名前
パスワード
ログイン
ログインしたままにする
パスワード再発行
新規登録
このサイトについて
みんなの自動翻訳@TexTra®とは
ヘルプ一覧
「みんなの自動翻訳@TexTra®」は、国立研究開発法人情報通信研究機構（NICT）が開発した自動翻訳サイトです。
最新の自動翻訳研究に基づく「高精度自動翻訳エンジン」が無料でご利用いただけます。
※利用登録に際しては、氏名・メールアドレスなどの個人情報は必要ありません。