

対訳データを集積し ニューラル翻訳のモデルを配布する エコシステム『**翻訳バンク**』

情報通信研究機構(NICT)
隅田英一郎

【第三部】 データ共有に向けて

16:50 – 17:50

(7) データ共有による研究とビジネスの共創 ～ A I オープンデータ協議会の発足と活動について～

田丸 健三郎 日本マイクロソフト株式会社 業務執行役員 NTO

深層学習の活用が様々な研究分野で広がる中、データアクセシビリティが新たな課題として注目されています。特にマルチモーダル、ヘテロなデータを必要とする場合、このデータアクセシビリティが大きな課題となっています。一定の額を支払う事により必ずアクセスできる音楽データと異なり、自然言語、気象、公共交通、地理といった様々なデータを複合的に使用する場合、個々個別に交渉、契約を結ばなくてはなりません。A I オープンデータ協議会は、データアクセシビリティ、AIサービス基盤の提供を通して社会課題の解決に取り組みます。

(8) 対訳データを集積しニューラル翻訳のモデルを配布する翻訳バンク

隅田 英一郎 情報通信研究機構 フェロー

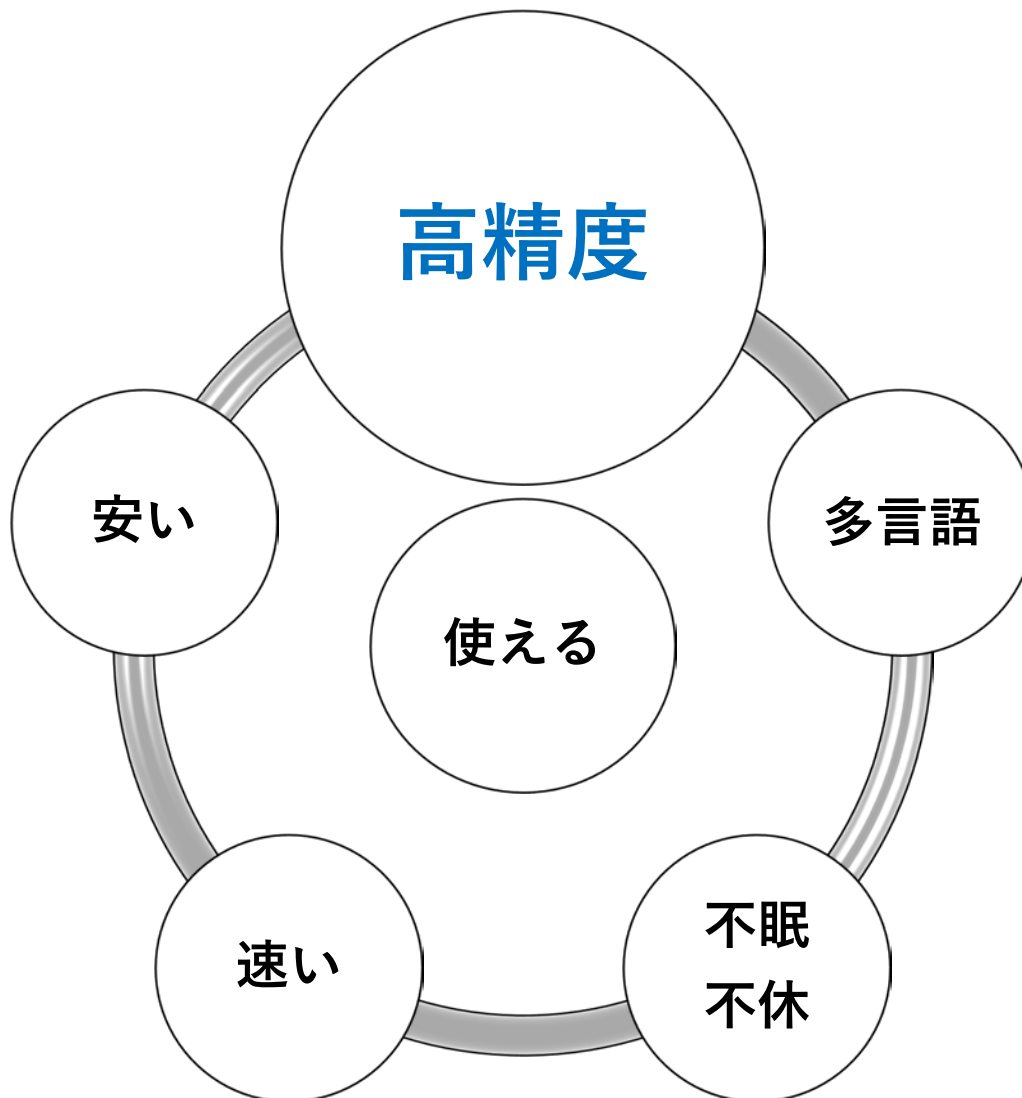
言語資源の有用性は広く認知されている。特に、対訳コーパスは高精度のニューラル翻訳を構築するための必須要素であり、量が多ければ多いほど高精度を実現できると考えられている。総務省とNICTは日本語と多言語の対訳コーパスを大規模に集積する新しい仕組（翻訳バンクと呼んでいる）を創出し運用している。翻訳バンクでは、対訳データを流通させるのではなく、対訳データの2次生成物であるニューラル翻訳のモデルを配布している。本講演では翻訳バンクの特徴とその進捗状況についてご紹介する。

目次

- ① 自動翻訳は昔と違い**使える！**
- ② 自動翻訳は**典型的なAI**である。
- ③ 翻訳バンクでデータの**蓄積・自動翻訳の高精度化が停まらない。**

① 日本語と
英語の自動翻
訳は使える！

① 日本語と英語の自動翻訳は**使える**！



自動翻訳の破壊的实力



みらい翻訳社長
栄藤稔氏

私たちが開発したAI翻訳の能力は、ビジネス会話の和文英訳なら**TOEICスコアで900点の人の能力を越えている。**

NMTで英作文が容易に →働き方改革や発信力増大



太洋物産
山本さん

顧客に出す提案書
や資料の英文作成
で徹夜や残業をし
なくて済むようにな
った。



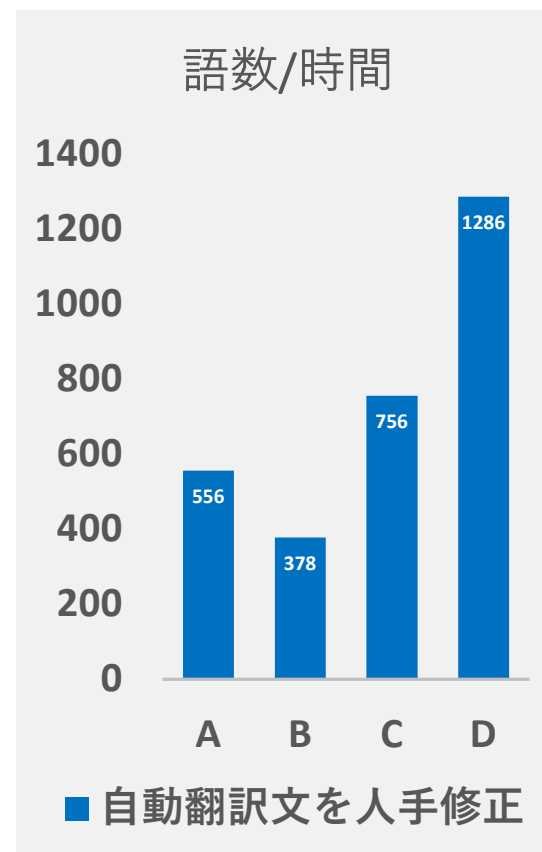
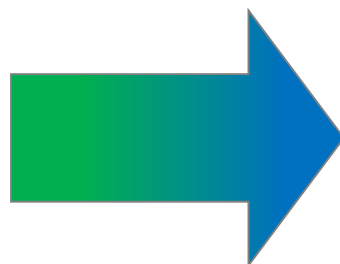
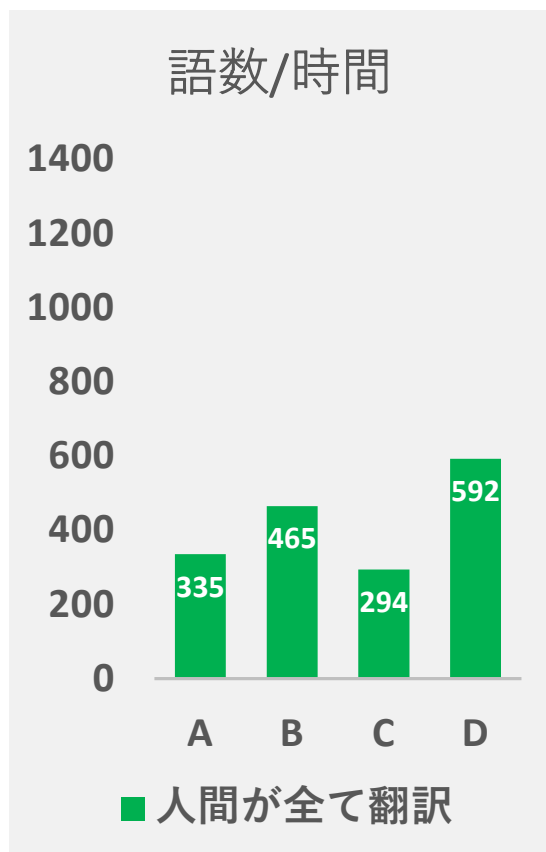
九州大学大学院
三浦研究室

修士の学生が、日本語
で論文を書き、その自
動翻訳結果を修正して
国際会議に論文を投稿
できるようになった。
しかも、採択された。

AERA2018年3月5日号 <https://dot.asahi.com/aera/2018030100036.html?page=2>

プロの翻訳者の効率も改善！

～人間 vs. (自動翻訳 + 人間)～



A、B、C、D、4人のベテラン翻訳者の協力の下、実験しました。

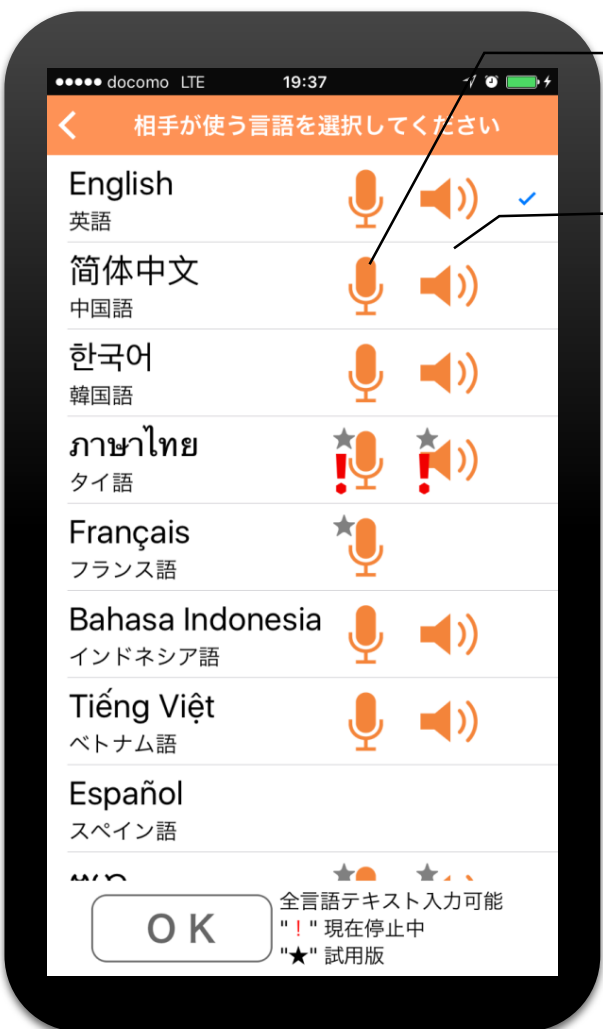
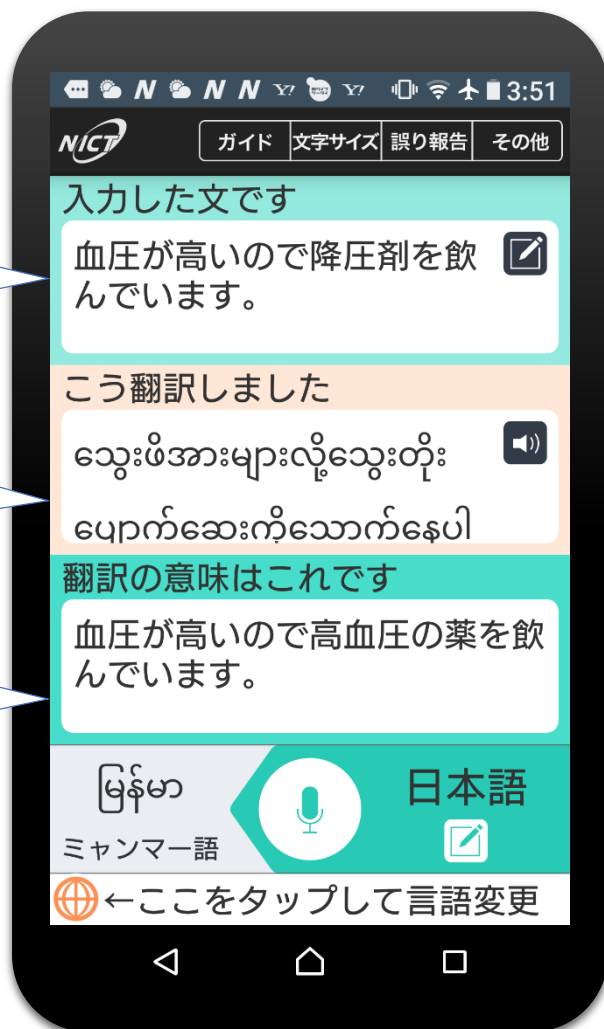
十印の三笠氏@JTF セミナー（2018/3/8）資料から改変

無料の.....をお試しく下さい・

入力した文

翻訳結果

結果の確認

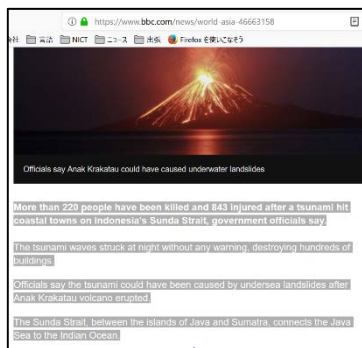


音声入力可
18言語

音声出力可
15言語

翻訳可能
31言語

ニューラル機械翻訳（NMT）を 「みんなの自動翻訳@TexTra」で試せる



- WEBベースのコピー
- OFFICEのアドインでファイル単位に
- CAT(MEMSOURCE,TRADOS,MemoQ)と連携
- APIでシステム構築

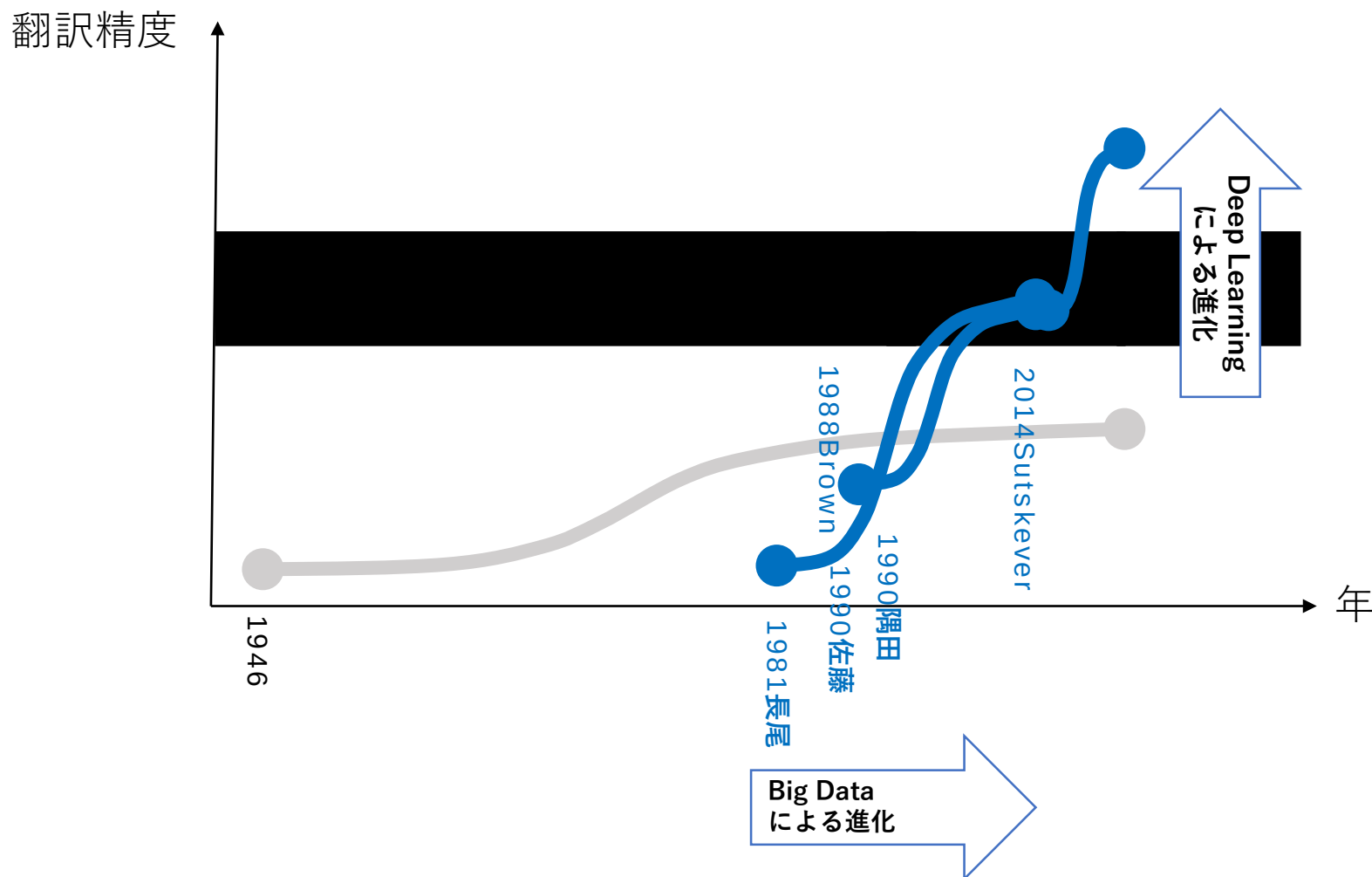
英語 → 日本語
↔
汎用NMT 【英語 - 日本語】 1
翻訳

More than 220 people have been killed and 843 injured after a tsunami hit coastal towns on Indonesia's Sunda Strait, government officials say. The tsunami waves struck at night without any warning, destroying hundreds of buildings. Officials say the tsunami could have been caused by undersea landslides after Anak Krakatau volcano erupted. The Sunda Strait, between the islands of Java and Sumatra, connects the Java Sea to the Indian Ocean.

インドネシアのスンダ海峡政府当局者によると、インドネシアの沿岸都市に津波が押し寄せた後、220人以上が死亡し、843人が負傷したという。津波は、何の警告もなく夜に襲ってきて、何100もの建物を破壊しました。アナク・クラカタウ火山噴火後の海底地滑りで津波が発生した可能性があるという。スンダ海峡は、Javaとスマトラの間にあり、ジャワ海をインド洋に接続している。

② 自動翻訳
は典型的なAI
である。

自動翻訳の歴史



対訳 (Big Data) から深層学習で 翻訳過程をニューラルネットに したニューラル機械翻訳 (NMT)

Big Data

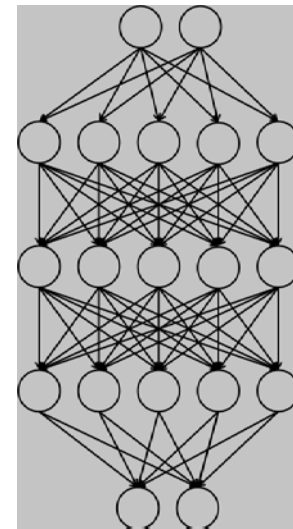
人類は、国家や地域、民族や世代など、あらゆる境界を越えて、相互の理解を深め、知恵を交わすなかで、発展してきました。コミュニケーションは人類社会を支えるもっとも重要な活動であり、情報通信技術はそのコミュニケーションを支える基礎であります。情報通信技術はまた、人類の高度な知的活動と経済活動を支える基盤でもあります。

Humanity has achieved progress as it has deepened its mutual understanding and shared its wisdom, overcoming barriers due to national, regional, ethnic, generational, and other differences. Communication is the most critical activity in human society, and information and communications technology (ICT) is the basis of that communication. ICT is also the infrastructure that supports humanity's advanced intellectual and economic activities.

<http://www.nict.go.jp/>および<http://www.nict.go.jp/en/index.html>

深層学習

ニューラルネット



進化する機械学習パラダイス ～改正著作権法が日本のAI開発をさらに加速する～ <https://storialaw.jp/blog/4936>

第三十条の四

著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には、**その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。**ただし、当該著作物の種類及び用途並びに当該利用の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

(中略)

二 **情報解析**（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、映像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう。第四十七条の五第一項第二号において同じ。）の用に供する場合

(以下略)

柿沼 太一 弁護士

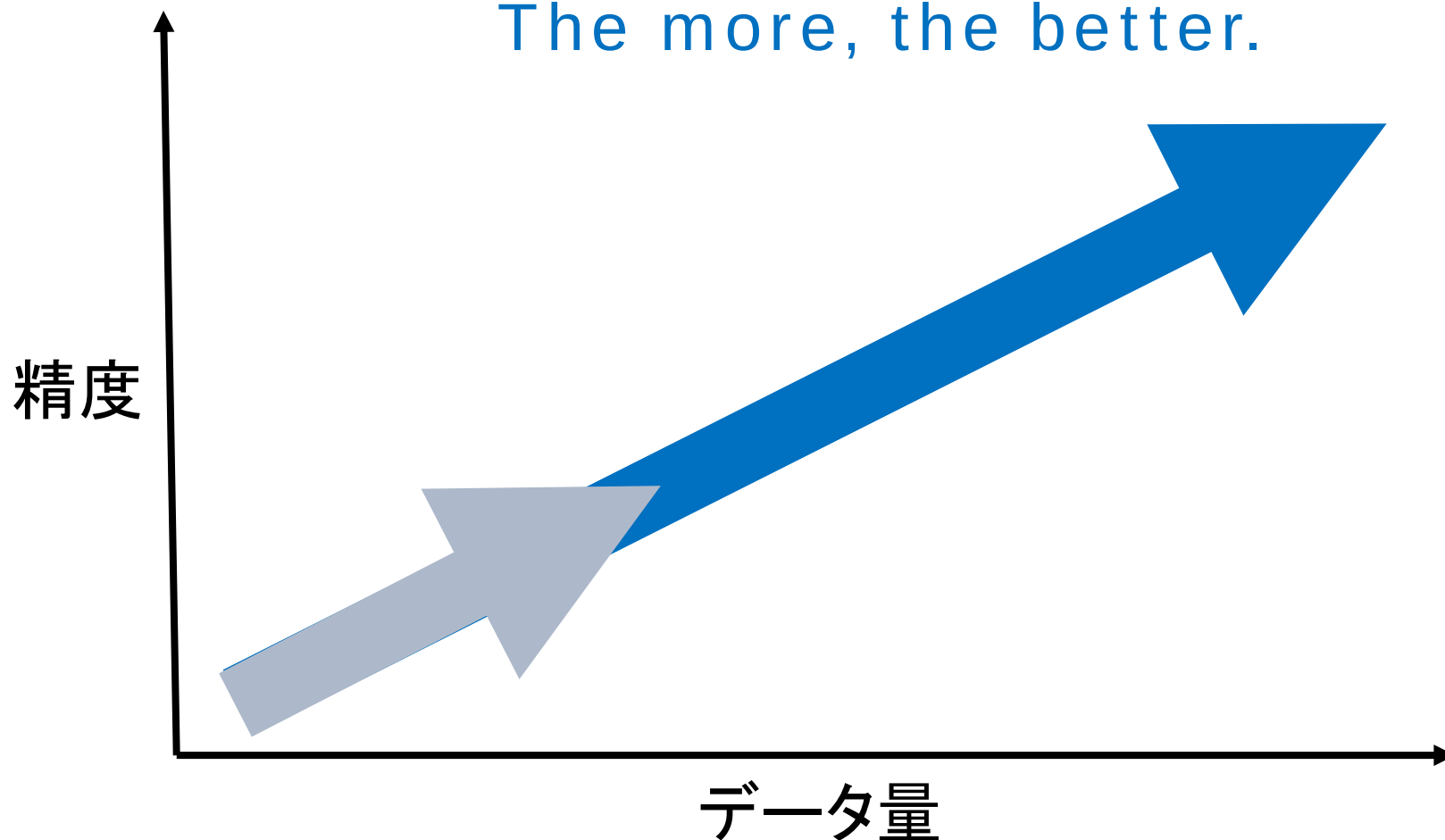


<https://storialaw.jp/lawyer/3041>

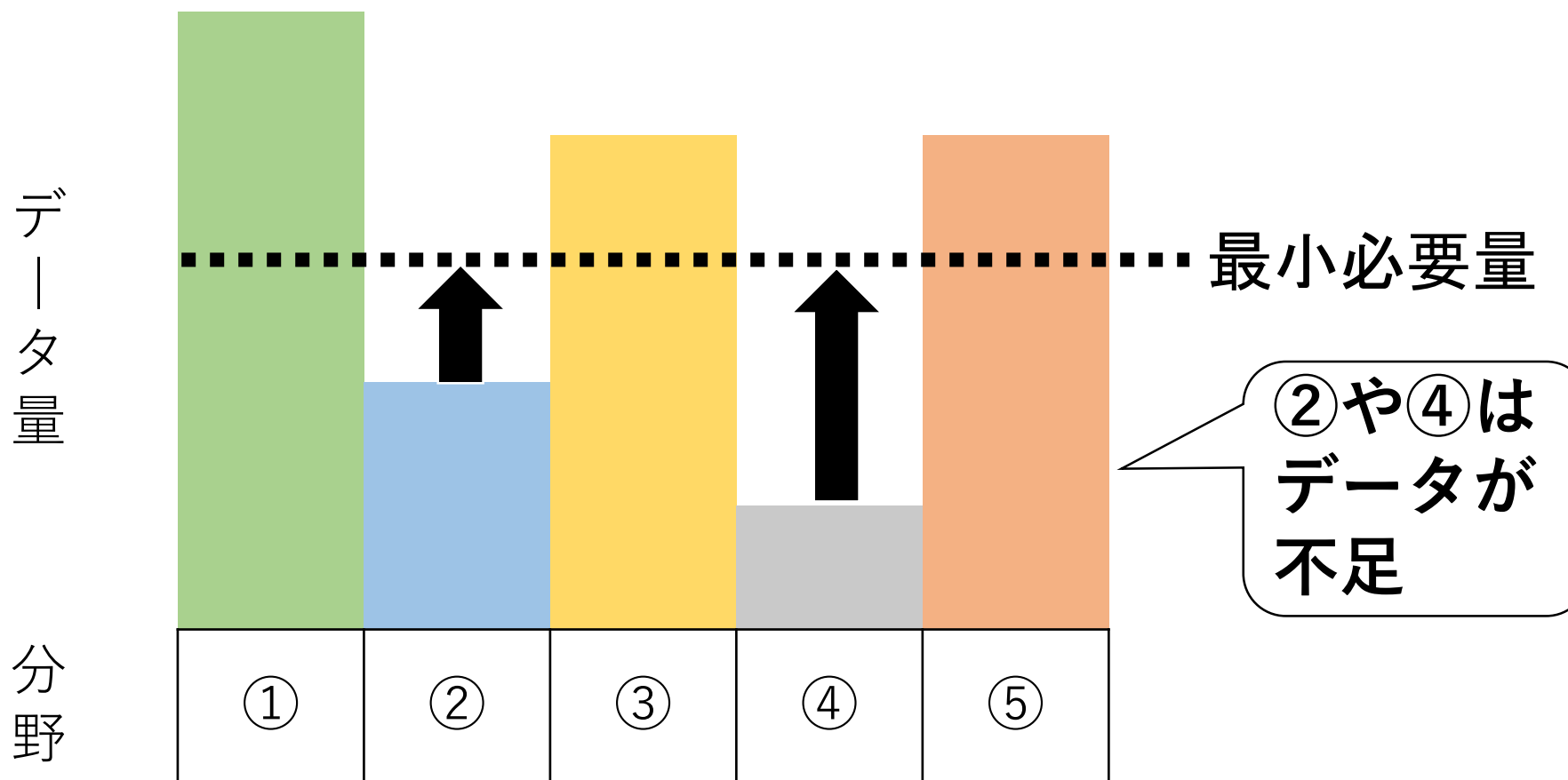
③ 翻訳バンクで
データの蓄積・自
動翻訳の高精度化
が停まらない。

経験則： 精度 $\propto$ データ量

The more, the better.



データ不足の分野の解消が課題

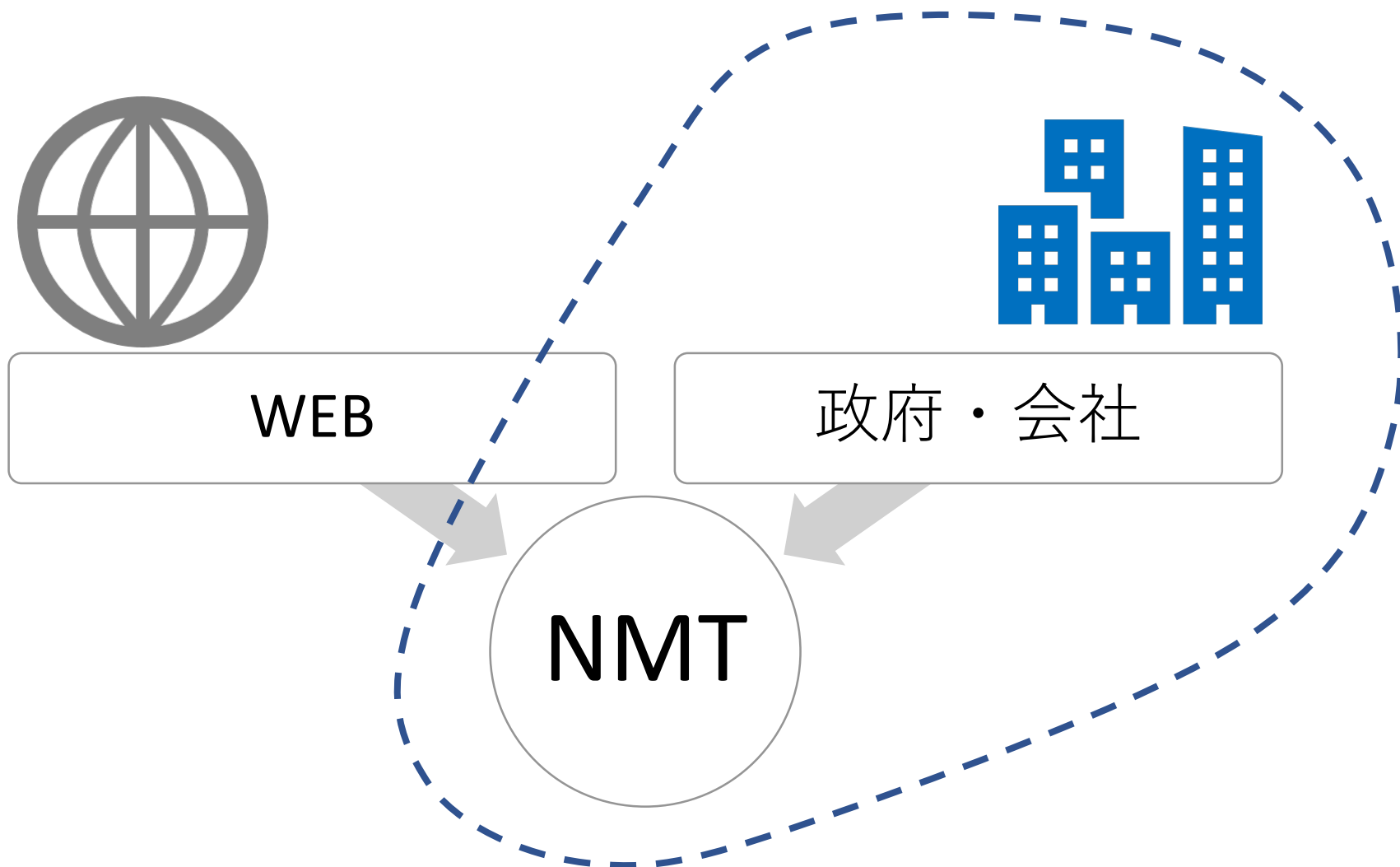


最先端の自動翻訳の訳文も色々

日本語原文	二つのNMTの訳文	品質
電気がつきません	There is no electricity.	×
	The light doesn't work.	○
トイレが流れません	The toilet doesn't flow.	×
	The toilet doesn't flush.	○
地図上に行き方を書いてもらえませんか？	Can you write how to go on the map?	△
	Could you please write down how to get there on the map?	○

<https://style.nikkei.com/article/DGXMZO19312810X20C17A7000000?channel=DF260120166490>

信頼度の高いデータを大量に



翻訳データ (原文とその訳文の対)



経済産業省
Ministry of Economy, Trade and Industry

文字サイズ変更



検索

[ホーム](#)
[経済産業省について](#)
[お知らせ](#)
[政策について](#)
[統計](#)
[申請・お問合せ](#)


[お知らせ](#)
[ニュースリリース](#)
[2014年度一覧](#)
[特許庁と独立行政法人情報通信研究機構（NICT）は特許文献の機械翻訳に関](#)
[意しました](#)


印刷

特許庁と独立行政法人情報通信研究機構（NICT）は特許文献の機械翻訳に関する協力に合意しました

本件の概要

特許庁と独立行政法人情報通信研究機構（NICT）は、外国語特許文献の機械翻訳の必要性の高まりを受け、中国語、ASEAN言語等の機械翻訳の精度向上及び活用促進のための協力を進めることに合意しました。これにより、民間の特許文献機械翻訳サービスの高品質化や、特許庁での機械翻訳活用による特許審査の効率化、高品質化などが期待されます。

お知らせ

- 会見・スピーチ・談話
- ニュースリリース
 - 2016年度一覧
 - 2015年度一覧
 - 2014年度一覧
 - 2013年度一覧

翻訳データを公的機関に集積

- 日本では、年間**2000**億円以上が翻訳に投じられている。
- 5億文の翻訳が生産されていると推定される。
- 当初目的に以外に使われることなく死蔵される。
- 自動翻訳システム高精度化に再利用すれば、新たな価値を生む。

The screenshot shows the official website of the Ministry of Internal Affairs and Communications (MIC) of Japan. The page is titled "『翻訳バンク』の運用開始" (Start of Operation of 'Translation Bank'). It features a large blue dashed oval graphic with the text "『翻訳バンク』の運用開始" and "—自動翻訳システムのさらなる高精度化に向けて様々な分野の翻訳データを集積—". The page includes a breadcrumb trail: "総務省トップ > 広報・報道 > 報道資料一覧 > 『翻訳バンク』の運用開始". The date "平成29年9月8日" (September 8, 2017) is displayed. The text explains that the MIC and NICT are starting the operation of the 'Translation Bank' to collect translation data from various fields to improve the accuracy of automatic translation systems. It also lists key points: starting the collection of translation data from various fields and utilizing neural machine translation for higher accuracy.

総務省
MIC
Ministry of Internal Affairs
and Communications

ご意見・ご提案 English

サイト内 関連サイト

Google カスタム検索

総務省トップ > 広報・報道 > 報道資料一覧 > 『翻訳バンク』の運用開始

報道資料

平成29年9月8日

総務省
国立研究開発法人情報通信研究機構

『翻訳バンク』の運用開始

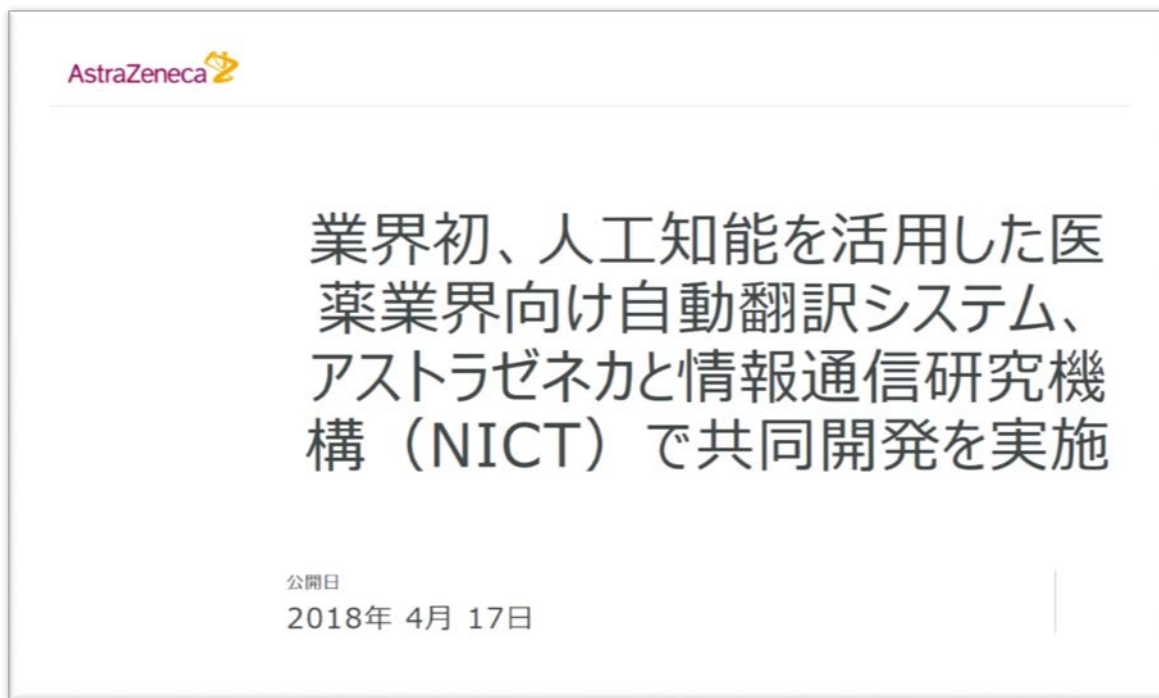
—自動翻訳システムのさらなる高精度化に向けて様々な分野の翻訳データを集積—

総務省と国立研究開発法人情報通信研究機構(NICT、理事長: 徳田 英幸)は、自動翻訳システムの様々な分野への対応や高精度化を進めるため、オール・ジャパン体制で翻訳データを集積する『翻訳バンク』の運用を開始します。これにより、社会・経済活動のグローバル化が進む中で我が国の国際競争力の強化に貢献します。

【ポイント】

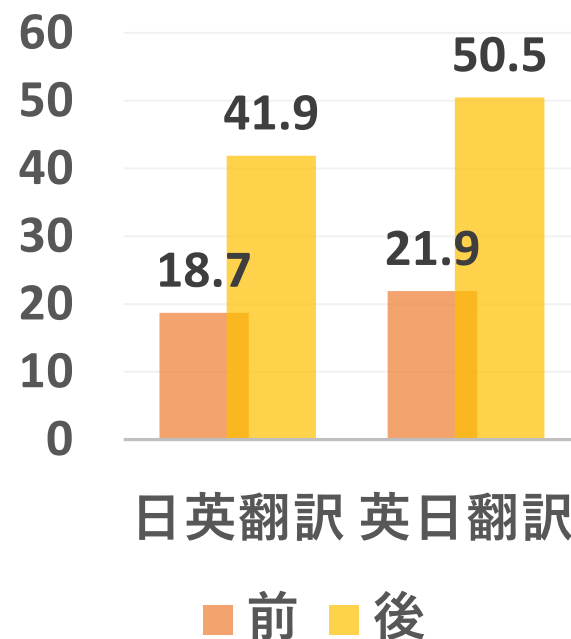
- オール・ジャパンで様々な分野の翻訳データを集積する『翻訳バンク』の運用を開始
- ニューラル機械翻訳での活用等に向けて翻訳データを集積し、多分野対応や高精度化を実現

先進事例～製薬会社アストラゼネカ～



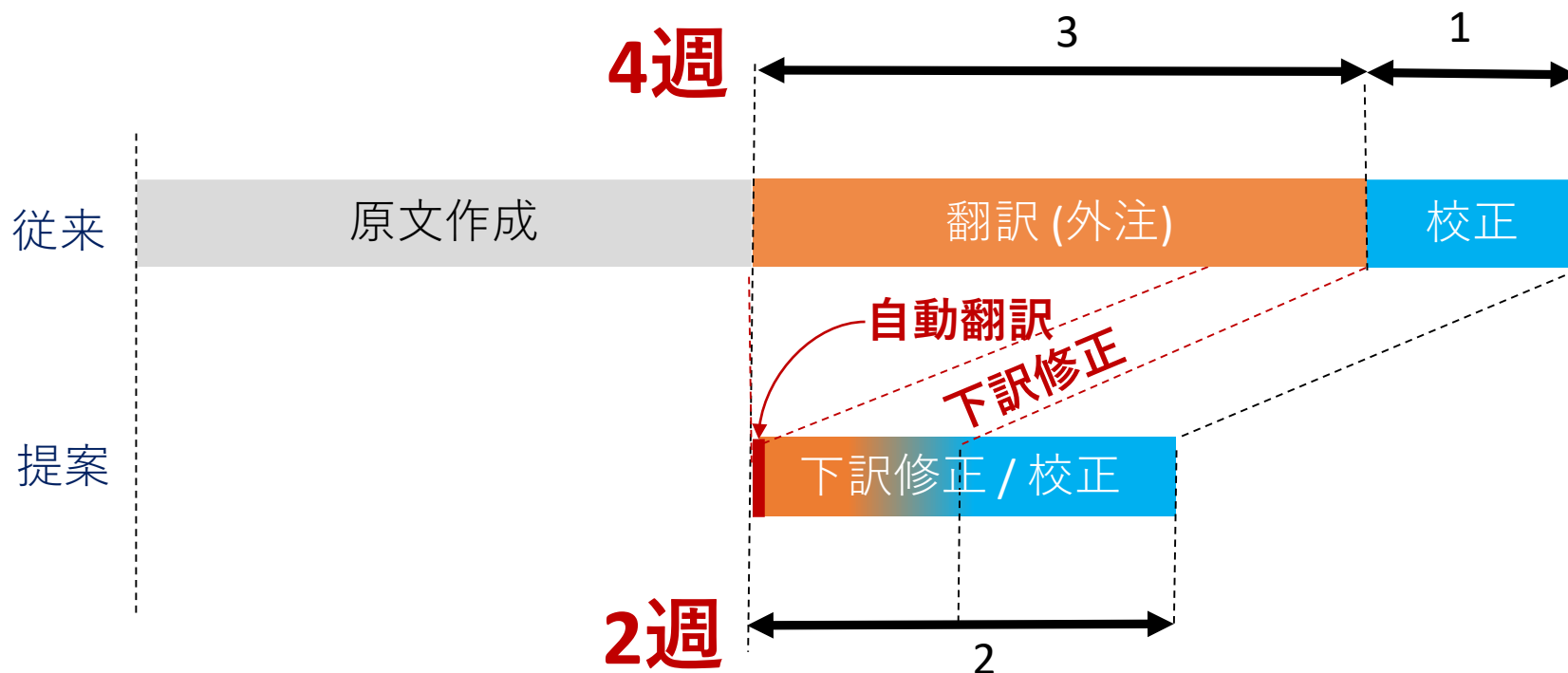
<https://www.astrazeneca.co.jp/media/press-releases1/2018/2018041701.html>

BLEU



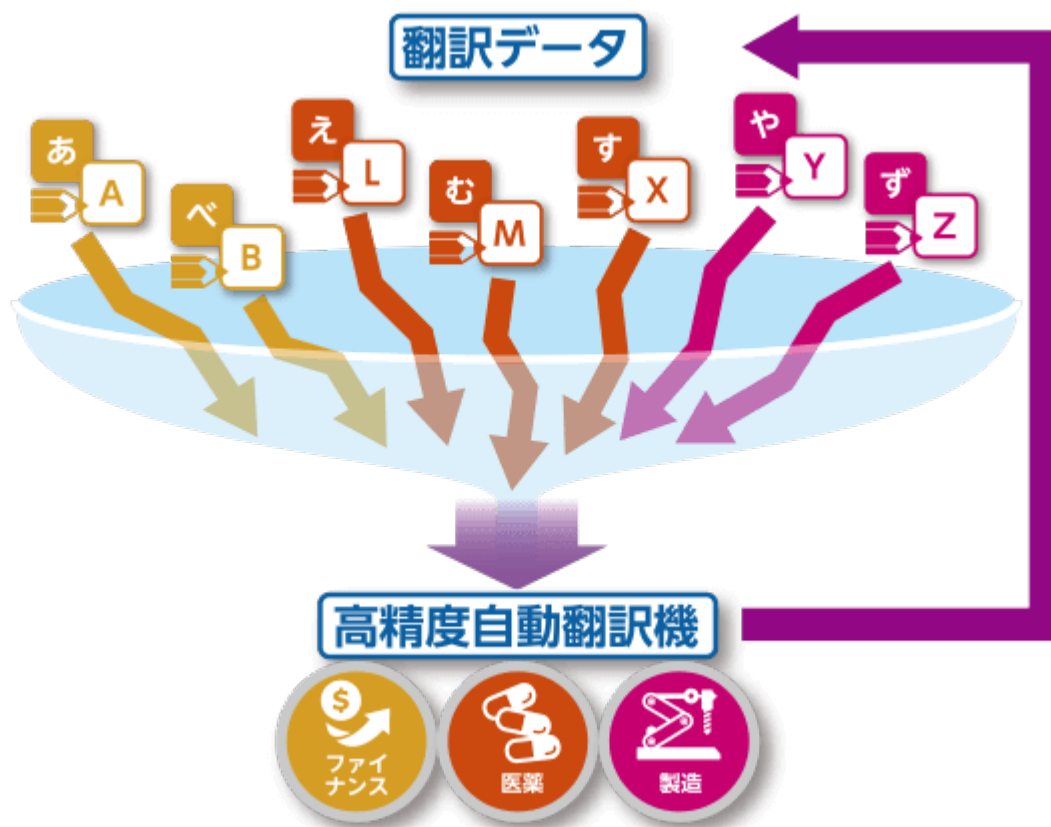
※ BLEUとは、複数のプロ翻訳者が予め作成した訳文と自動翻訳の訳文との類似度であり、自動翻訳の品質評価のため広く利用される。類似度が大きいほど良い訳と判断される。

治験実施計画書の作成期間を半減

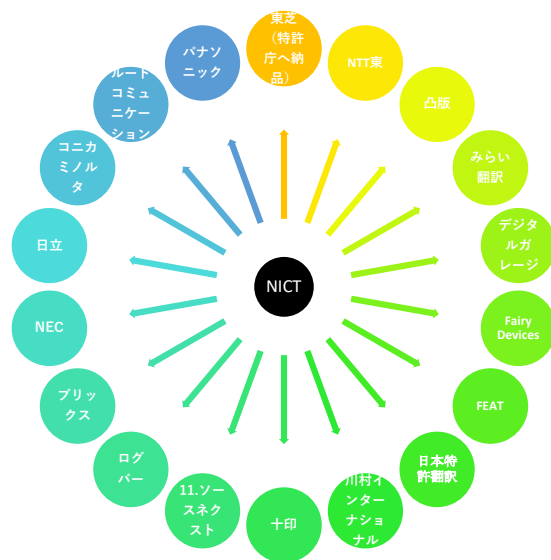


- 新薬が早く患者に届く。
- 新薬を世に出す費用が安くなる。

散在する対訳コーパスを 一か所に集める



ニューラル機械翻訳（NMT）を 各社からビジネスに導入可能

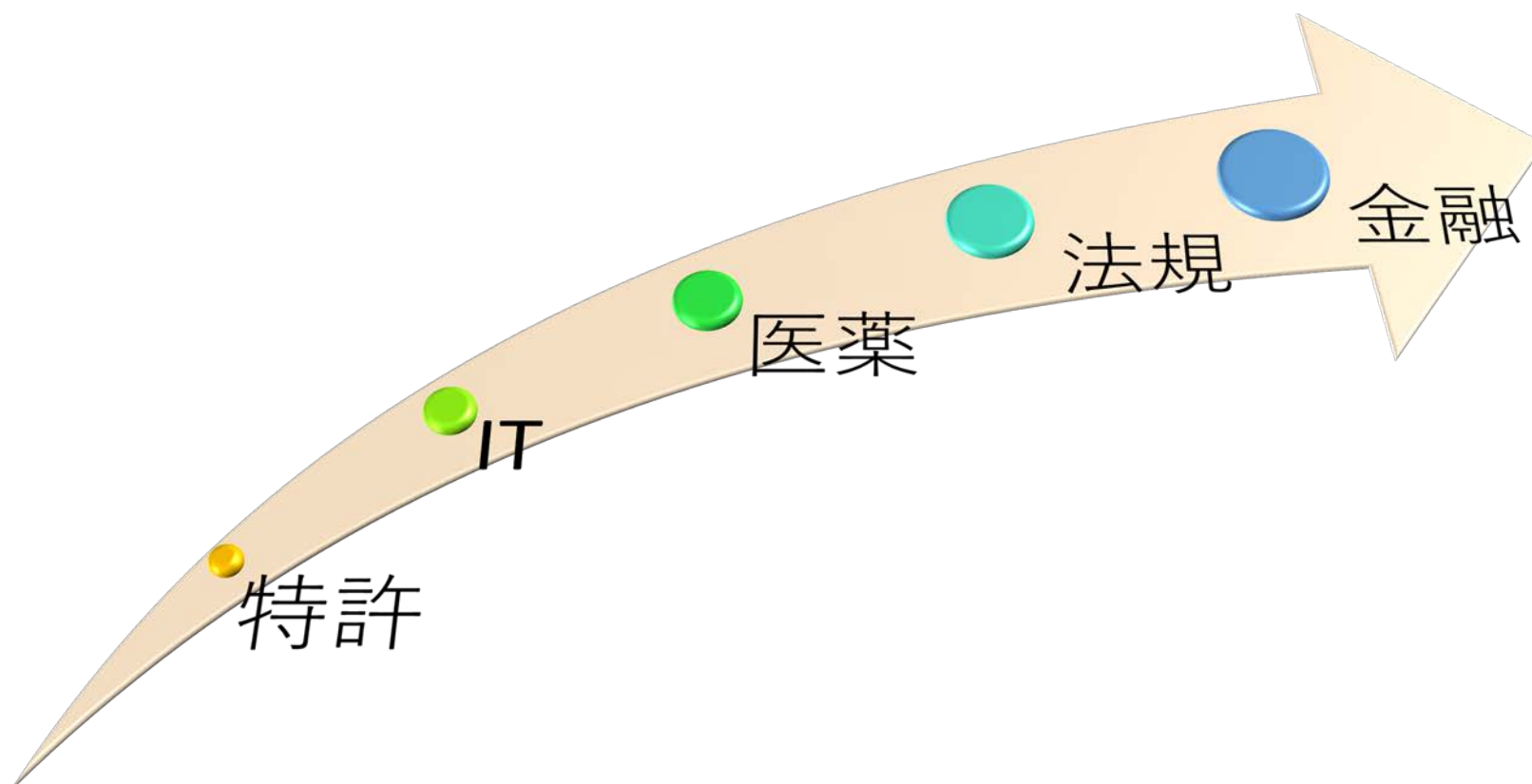


1. NEC
2. 日立
3. 東芝
4. 富士通
5. パナソニック
6. コニカミノルタ
7. NTT東
8. 凸版
9. リクルートコミュニケーションズ
10. デジタルガレージ

11. ソースネクスト
12. ログバー
13. Fairy Devices
14. FEAT
15. MindWord
16. 日本特許翻訳
17. みらい翻訳

18. 川村インターナショナル
19. 十印

翻訳バンクの次のターゲットは!?



ポジティブ・スパイラル ～『翻訳バンク』という名のエコシステム～



時間	次第
15:00 - 15:05	主催者挨拶 総務省
15:05 - 15:40	基調講演 「データがAIの燃料」となる時代の研究開発戦略 大学共同利用機関法人情報・システム研究機構国立情報学研究所 所長 喜連川 優
15:40 - 15:55	翻訳バンクの概要説明 NICT フェロー 隅田 英一郎
15:55 - 16:20	講演「医薬品開発向けAI翻訳の評価と導入」 アストラゼネカ株式会社 執行役員・サイエンス&データテクノロジー統括部長 田中 倫夫
16:20 - 16:55	講演「著作権法改正がAI開発に与える衝撃」 STORIA法律事務所 弁護士 柿沼太一
16:55 - 17:55	パネルディスカッション
17:55 - 18:00	閉会挨拶 NICT 理事長 徳田 英幸

シンポジウムに
是非ご参加くだ
さい

3月6日@日本橋