機械翻訳と社会実装

追手門学院大学　心理学部心理学科
人工知能・認知科学専攻
井佐原　均

Yes or No

- 人工知能型機械翻訳は
  人間を超えるか
  ？

- 人工知能型機械翻訳は
  普通の人間が一つの文を翻訳する能力を超えるか
  Yes

- 人工知能~~型機械翻訳~~は
  人間を超えるか
  No

機械翻訳の現状

- 2016年後半に統計翻訳（や規則翻訳）からニューラル翻訳への移行
  - 深層学習・ニューラルネットを用いた人工知能型の機械翻訳システム
  - 単語の置き換えから、文から文への置き換え（翻訳）へ。
  - Google や Microsoft の発表で話題に。
  - 一般的な人よりも翻訳能力が高い。
    - ✓流暢度は完ぺき。忠実度も人間並み。

統計翻訳とニューラル翻訳

原文：可燃ごみは、週2回、決められた曜日に収集します。

統計翻訳：Burnable waste is collected in the day of the week decided twice a week.

ニューラル翻訳：Combustible waste is collected twice a week on designated days of the week.

原文：少枝・葉は、30センチメートル程度の長さにしてから、2から3束づつ出してください。

統計翻訳：Please take each 3 bunches out of 2 after the small branch and the leaf are made about 30 centimeters of length.

ニューラル翻訳：Make the twigs and leaves about 30 cm long, and then put out 2 to 3 bundles at a time.

コンピュータと翻訳（自然言語処理技術）

- コンピュータの出現と同時に研究が始まる
  - 第2次世界大戦が終わったころ（1945年）
  - 機密を伝達するための暗号の研究

- 暗号：言葉　⇒　意味不明の文字列　⇒　言葉
- 翻訳：日本語　⇒　　英語　　⇒　日本語

- 言葉を「理解」する

機械翻訳の歴史

- １９５０年頃に研究が始まる
- １９８０年代は日本が世界のトップ
  - 多くの企業の参入、国際会議での発信
- **日本の（機械翻訳ビジネスの）衰退**
  - ➢当時の機械翻訳の精度は低かった
    - ➢特に日本語と他言語の間
  - **➢機械翻訳を翻訳サービスの一環と考える視点の欠如**
- １９９０年頃からデータを利用した機械翻訳の発展
  - 用例翻訳システム、統計翻訳システム
  - 大量の言語データの存在とコンピュータの性能向上
- **ニューラル機械翻訳による大幅な精度向上**

機械翻訳技術の発展

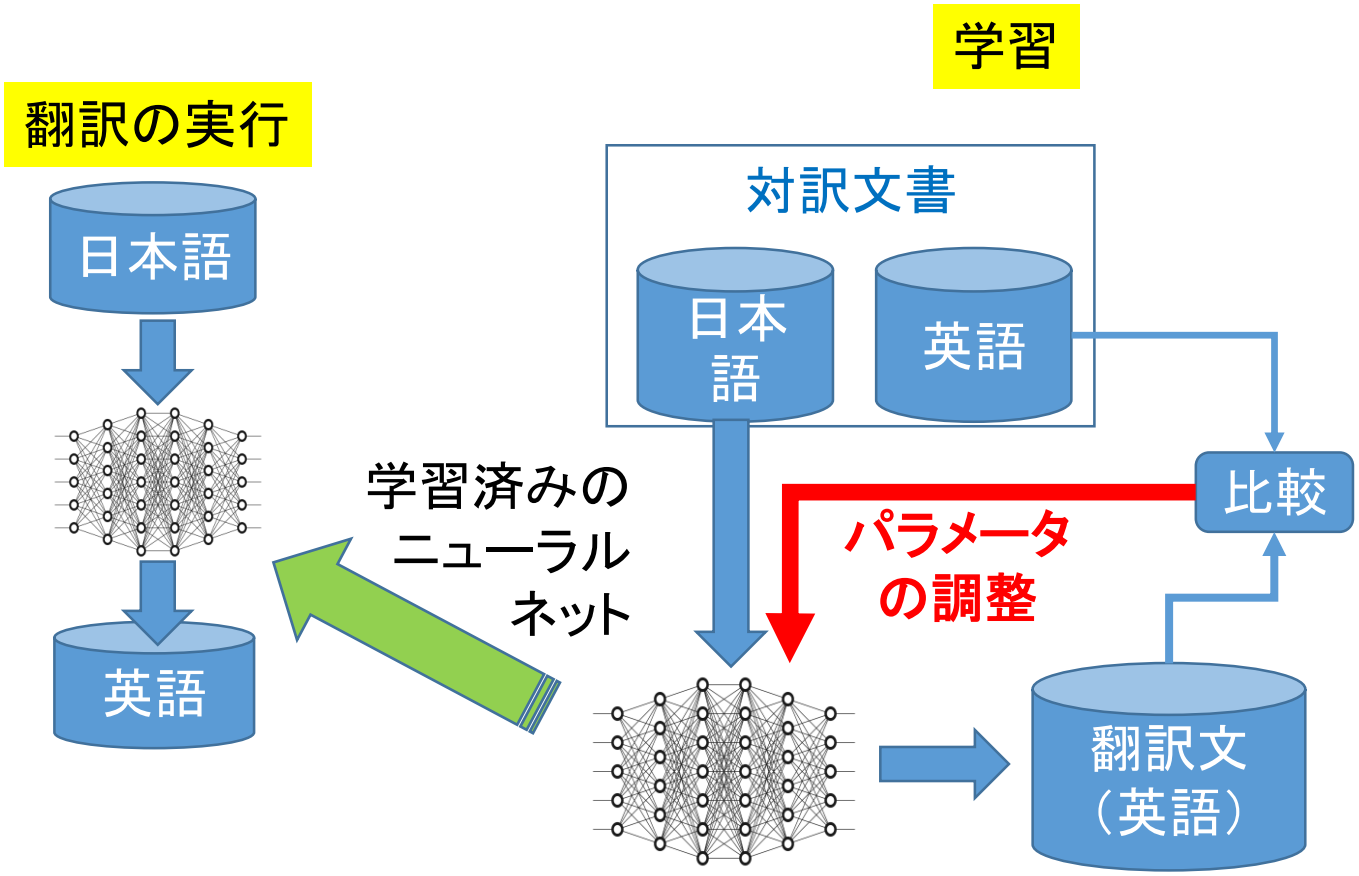
- 規則翻訳
- 統計翻訳
- 用例翻訳
- ニューラル翻訳

機械翻訳技術の発展

- 規則を用いた機械翻訳（小規模なコーパスを知識源として利用）Rule-based Machine Translation (RBMT)
  - コンピュータに文法規則と辞書を覚えさせる。
  - 細かなチューニングが可能。
  - 長い研究開発成果の蓄積がある。
  - 開発・管理・改良にコストがかかる。
- 統計翻訳（大規模な対訳コーパスを統計量を得るために利用）Statistic Machine Translation (SMT)
  - 対訳データからの自動学習によるシステム。
  - 異なるタイプの言語間では学習用のデータ（対訳テキスト）の量が膨大になる
- 用例翻訳（大規模な対訳コーパスを用例として利用）
  - Example-based Machine Translation (EBMT)

ニューラル翻訳

- 人工知能型機械翻訳
  - ニューラル機械翻訳
  - Neural Machine Translation
  - NMT



ニューラル翻訳（利点と限界）

- 逗子市小坪5-1の小坪海岸トンネル鎌倉側で、9月24日0時頃、大きな崖崩れが発生しました。そのため、小坪海岸トンネルや国道134号線における通行に規制がかかっています。

➢（統計翻訳）Tunnel Kamakura, zushi, kotsubo 5-1 kotsubo coast as 9/24 about a big landslide has occurred. As a result, traffic in kotsubo coast tunnel and route 134, on regulations.

➢（ニューラル翻訳）A large cliff collapse occurred around 0 o'clock on September 24th at the Kobosa coast tunnel Kamakura side of Zushi-shi Kobosa 5-1. For that reason, traffic restrictions are imposed on the Kosugato coast tunnel and National Route 134.

ニューラル翻訳（絶え間ない性能向上）

- 逗子市小坪5-1の小坪海岸トンネル鎌倉側で、9月24日0時頃、大きな崖崩れが発生しました。そのため、小坪海岸トンネルや国道134号線における通行に規制がかかっています。
- A large cliff collapse occurred around 0 o'clock on September 24th at the Kobosa coast tunnel Kamakura side of Zushi-shi Kobosa 5-1. For that reason, traffic restrictions are imposed on the Kosugato coast tunnel and National Route 134.
- A large landslide occurred around midnight on September 24th at the Kamakura side of the Kotsubo coast tunnel at 5-1 Kotsubo, Zushi City. Therefore, traffic on the Kotsubo Kaigan Tunnel and National Highway No. 134 is restricted.

ニューラル翻訳（利点と限界）

- 翻訳精度は飛躍的に向上した。

- でも。。。。

- 訳の抜けや誤りが発生することがある
- **低頻度の語句**については、失敗が多い
- 同じ単語を繰り返し出力することがある

✓1つの文を1つの文に置き換える

✓前後の文の情報は使わない

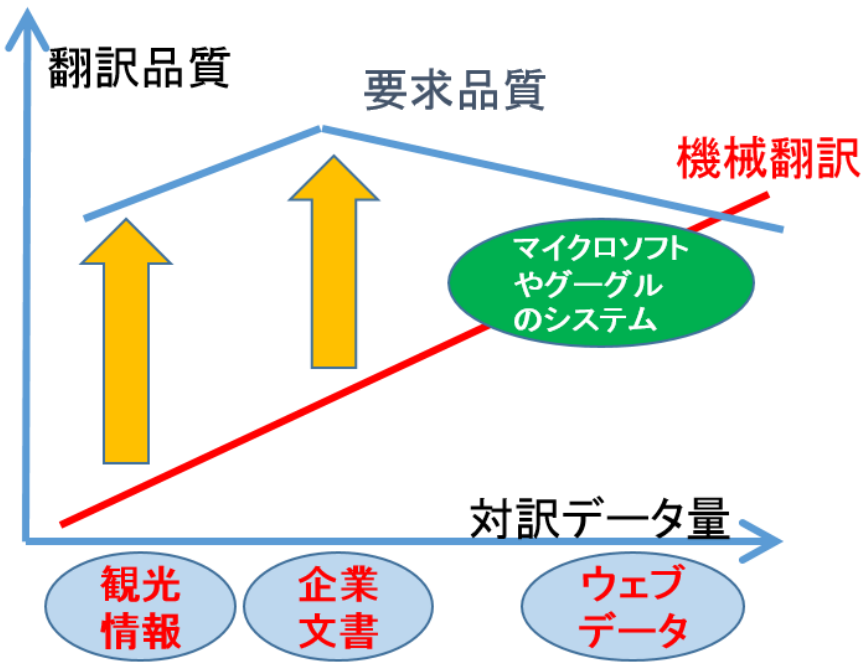
データと機械翻訳

- 規則翻訳
- データ型翻訳
  - 統計翻訳（単語レベル）とニューラル翻訳（文レベル）
  - データの質と量が翻訳品質に直結

機械翻訳の性能

- データ量と翻訳精度

- 翻訳の方向
  - 情報発信と情報受信
  - 日本語



課題を解決し、実用的な翻訳を実現

大規模テキストデータ（コーパス）と機械学習を用いたニューラル機械翻訳の実現

↓

Microsoft, Googleなど、機械翻訳サービスの品質が大幅に向上

平均的な人よりも翻訳能力は高い

低頻度の語句については、訳の誤りや抜けが発生することがある

↓

情報発信に活用できる機械翻訳システムを実現するには

観光、医療、自動車など
分野ごとの文書量は多くない

→ 個々の分野での**学習**が必要。

↓

用語の対訳辞書では不十分。長めの語句の対訳が必要。
重要語のみでなく、**典型的な句を取り出す**技術が必要。

改善法の例

抽出例
「シート エアバッグ スクイブ 回路 の 短絡 の 点検」
「P-RNAVに必要な機器要件を満足しなくなった場合」

→ 実社会への適用

機械翻訳の社会実装

- 東京都港区で実証実験。
- ビジネス翻訳での実証を計画中。

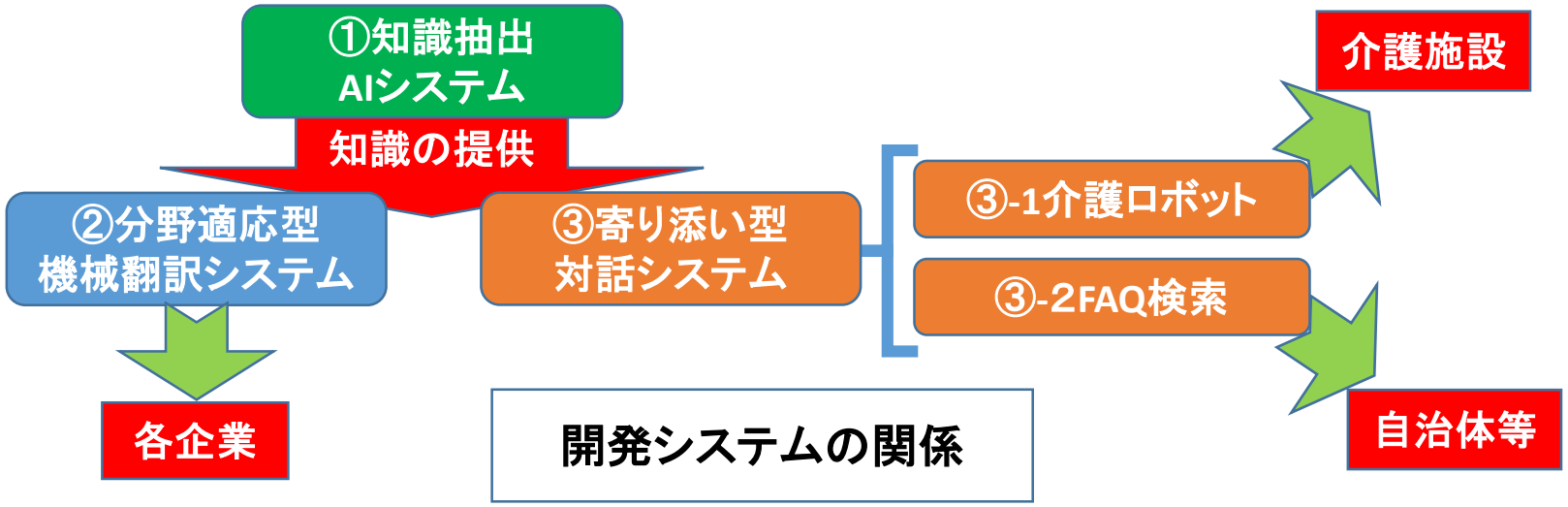
入力文の前変換、固有名詞・専門用語辞書の利用、分野ごとに学習したシステム

複雑な表現、固有名詞、住所表記などに対して、顕著な性能改善。

各総合支所区民課窓口サービス係（芝地区は区民課証明交付担当）及び台場分室		
総合支所	Regional City Office	comprehensive branch office
区民課	Residents Support Section	Inhabitant （訳抜けあり）
窓口サービス係	Counter Services Subsection	window contact service staff
芝地区	Shiba District	shiba district
台場分室	Daiba Annex	Daiba branch office
		Google翻訳

機械翻訳の社会実装

- 知の拠点あいち重点研究プロジェクト
  - 分野適応技術による自然言語処理技術のビジネス展開



機械翻訳の社会実装

- 翻訳実験用インタフェース（翻訳環境）を開発
  - 英日・日英、用語辞書の有無、に対応。
- 社内文書翻訳
  - 対象文書の対訳化（学習データ化）⇒翻訳実験。
- 特許翻訳
  - 大量の対訳を用いた翻訳実験
- ウェブページ翻訳
  - コロナ対応のウェブページ
  - 日本語ページの修正・公開⇒英語に自動翻訳⇒チェッカーによる修正⇒英語版の公開
- 観光情報翻訳
- 評価
  - 忠実度だけでなく、有効な文の割合や、ユーザの判断も検討。

機械翻訳に難しいこと

文の翻訳から
文書の翻訳へ

→

相手に応じた文
書の作成

伝えたい情報（意図）を伝える翻訳

「豆の鮮度にこだわるコーヒー」

“Coffees brewed with fresh coffee beans”

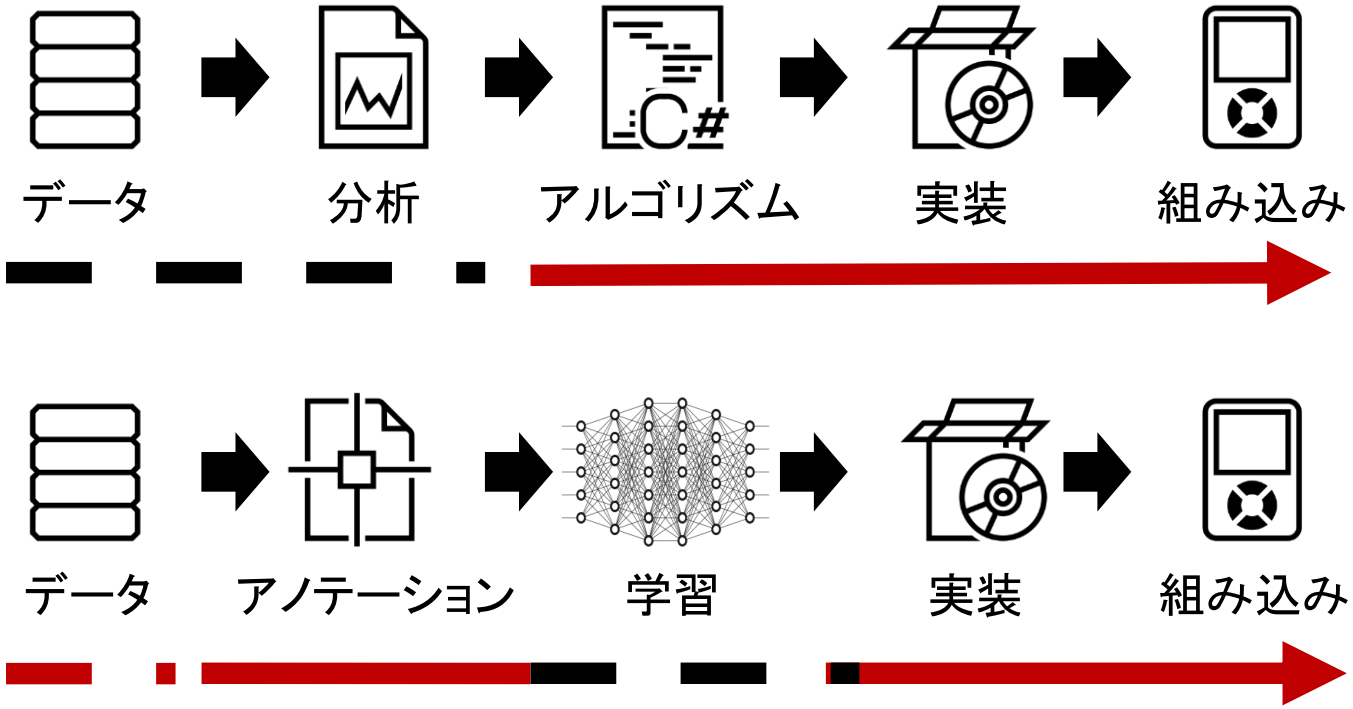
- 店主の「こだわり」を外国人に伝えるにはどうすればよいか。

人工知能の課題

- 現在の人工知能システムは大量の正解データを用いた学習によって精度を向上する。
- 正解データ（高品質のデータ）が必要
- 入力と出力を繋ぐ部分はブラックボックス

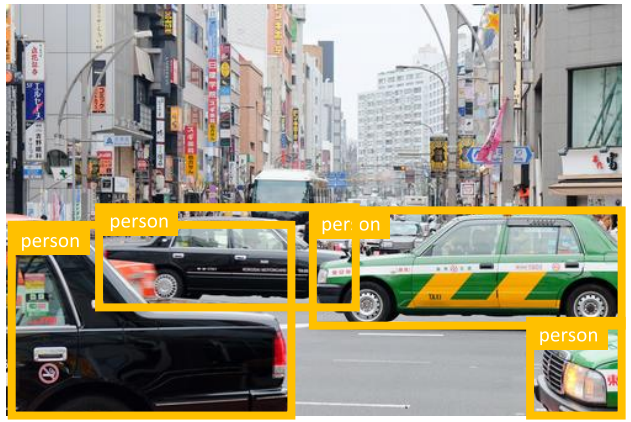
- 理由を説明できる人工知能が必要
- アルゴリズムだけではなく、データやアノテーションを保証

人工知能の実装と責任範囲



データの保証

- 情報付与の「誤り」と「悪意」



言葉と脳

- 人間が大量の言語データからの統計量を記憶し、言語理解を行っているとは思いにくい。
- コンピュータは人間とは異なった解法

- 脳の制約による言語の成立
- 人間の脳の構造が（自然界の環境の中で）言葉（言語の体系）を生み出したのではないか。

- 人間の脳で処理しやすい言語体系。

人間と同じように教える（覚える）

- 第1言語（日本語）の獲得と第2言語（英語）の獲得法は異なる
- 言葉（英語）を学ぶときに
  - 文法
  - 辞書

第２言語獲得

- 言葉（日本語）を学ぶときに
  - 経験
  - 環境

第１言語獲得

- 初期の自然言語処理の研究はコンピュータに文法と辞書を教えようとした。
- 人間の知識（知的活動）を模擬したシステムの開発が必要。

さらなる高精度化に向けて

- 知識と用例の融合
  - 大規模な言語資源を用いた知識処理
  - 機械翻訳・検索等の実用システムの開発

- 人間の知識（知的活動）を模擬したシステム
  - 人と同じように理解する（誤解する）
  - 人間は言語を見聞きしてから理解処理を開始しているのではなく、予測していた情報と言語の解析結果（途中結果）と照合している？？