第10回産業日本語研究会・シンポジウム
人工知能時代の産業日本語
～分かりやすい日本語の実現に向けて～

10年を迎えた産業日本語

産業日本語研究会 世話人会 代表／
豊橋技術科学大学 情報メディア基盤センター長・教授
井佐原　均

産業日本語（Technical Japanese）

- 「産業・技術情報を人に理解しやすく、かつ、コンピュータ（機械）にも処理しやすく表現するための日本語」
- 言語処理技術を活用することによって、明瞭な日本語文を作成することや高品質な翻訳文を低コストで作成すること
- 産業日本語研究会は自然言語処理や言語学の研究者、産業・技術情報を持つ様々な産業、言語サービス・言語ビジネスなど、広く日本語に関わる人々が、このような目標に向けて集い、議論し、研究・開発・普及活動を推進することを目指して設立されました。

Short History

- 2007年:「特許版・明晰日本語策定委員会」
  - 日本特許情報機構（Japio）による「技術用日本語」、「明晰日本語」という活動
- 2009年:産業日本語研究会の発足
  - 名称を「産業日本語」に統一（2008年）。
  - 特許版・産業日本語委員会を設置。
- 2016年：特許版・産業日本語委員会を産業日本語研究会に統合

１０年前と今

- 機械翻訳
  - 統計翻訳や用例翻訳といったデータに基づく機械翻訳システムが実用化されていたが、翻訳精度は不十分
  - ニューラル機械翻訳が実用化されたことにより、翻訳の精度が大幅に向上
- 他の自然言語処理技術応用分野においても、深層学習を利用することにより高性能のサービスが可能に

統計翻訳とニューラル翻訳の比較

- 可燃ごみは、週2回、決められた曜日に収集します。
- Burnable waste is collected in the day of the week decided twice a week.
- Burnable garbage is collected twice a week on the designated days of the week.
- 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。
- Please take each 3 bunches out of 2 after the small branch and the leaf are made about 30 centimeters of length.
- Branches and leaves should be about 30 centimeters in length, please put out 2 to 3 bundles each.

データ(テキスト)の重要性

- 自然言語処理の手法
  - 規則による手法⇒統計や機械学習による手法⇒深層学習
  - 学習に基づく手法である限り、質の高いデータが大量に必要
- 大量データによる深層学習によるシステム
  - 処理の過程が分からないブラックボックス的なシステム
  - 正確性の保証には、分野と応用に合った質の高いデータが重要

人にも分かりやすい日本語

- 産業・技術情報を表現する⇒「人に理解しやすい」ということも重要
- 産業日本語研究会のもとにライティング分科会、文書作成支援分科会、特許文書分科会を設置
- コンピュータによる特許ライティング支援
  - ✓「言い換えルール」の抽出
  - ✓2013年に特許ライティングマニュアル（初版）
  - ✓内容の見直しを行い、7つのカテゴリー、27のルールに再構成し、併せて、例文の追加・修正を行った第2版を発行

⇒ポスターセッション

産業日本語シンポジウム

- 第7回シンポジウム
  - ニューラル機械翻訳前夜の2016年2月に開催
  - テーマ：人工知能と産業日本語の出会い～先進的グローバル・ビジネスへの展開と躍進～
  - 現在に続く人工知能応用技術への期待と展望を語る場
  - 東京大学（第6回までの開催場所）から丸ビルホールへ
  - ✓産業文書を実際に利活用する方々へのアウトリーチ
- 第8回シンポジウム
  - ニューラル機械翻訳の実サービスが開始された直後の2017年3月に開催
  - いち早くニューラル機械翻訳に関する講演を取り入れた。

第10回産業日本語シンポジウム

- テーマ：人工知能時代の産業日本語～分かりやすい日本語の実現に向けて～
- ビジネス活用には、コンピュータの訓練データとなる日本語の質と量の問題が重要
  - ◆日本語の質：ネットワーク時代において、相手に伝わりやすい文章とはどのようなものであるか（招待講演2件）

第10回産業日本語シンポジウム

- 人とコンピュータのコミュニケーションが様々な場面で出現
- 人間同士のような親密なコミュニケーションは難しい
  - 完璧で丁寧な受け答えだけではなぜ人間は満足できないのか。（⇒特別講演）
- データの質と量が精度に直結する機械学習
  - データ共有の重要性がますます高まる。（データ共有の動きを2つ紹介）

言葉と人工知能

- 人間に勝つ人工知能
  - チェス（1988年）、将棋（2013年）、囲碁（2016年）
  - 足し算、引き算
- 人間よりコミュニケーションが上手な人工知能？
  - 人と人とのコミュニケーション
  - 脳の制約による言語の成立
  - 人と同じように理解する
  - 不気味の谷

人間らしいコミュニケーション

- コミュニケーションの基盤
  - 会話参加者の関係性
  - 共有する知識や常識
- ゆらぎの重要性
  - 言語的ゆらぎ
    - 語の省略、単語選択の間違い、言い間違いなどによるゆらぎ
  - 感情・感覚のゆらぎ
    - 感情や感覚からくるコミュニケーションのゆらぎ
  - 対話におけるゆらぎ
    - 膨らみのある会話、冗談、独創的な連想などを許容し、紡いでいく会話

産業の言葉、個人の言葉

- 個人：「人間的な」対応、言語
- ビジネス：正確な命令、正確な言語
  - 機械に命令する、コンピュータから知識を得る。
  - 単一方向のコミュニケーション

◆人間とコンピュータの双方向のコミュニケーション

産業日本語

- コンピュータ処理（＋人間の理解）
- 再利用（加工性）
- 学習データ（正しい言語・情報）

- データの使い方
  - **大学における数理・データサイエンス教育の全国展開**
  - **基礎編、応用編それぞれ10コマのウェブ教材の作成と公開**

- ニューラル時代の正しい言語

正しい日本語？

- この先生き残るにはどうすれば良いですか
  - G：　How can I survive in the future?
- この先生き残るにはどうすれば良いですか。
  - G：　How can I survive this afternoon?
- この先生き残るにはどうすれば良いですか？
  - G：　How can I survive in the future?

- この先、生き残るにはどうすれば良いですか

正しい日本語？？

- あつあつだから気をつけてね
  - G：Please be careful as it is hot
- あつあつだから気をつけてね。
  - G：It's hot, so be careful.
- 熱々だから気をつけてね
  - G：Be careful, so be careful