【AAMT/Japio特許翻訳研究会活動報告】

特許機械翻訳の課題解決に向けた機械翻訳技術解説

須藤 克仁

AAMT/Japio特許翻訳研究会 副委員長

（奈良先端科学技術大学院大学）

はじめに

- 本発表は研究会メンバーによる下記解説論文を紹介するものです
  - 共著者の皆様をはじめ、当研究会関係者に御礼申し上げます

今村賢治、越前谷博、江原暉将、後藤功雄、須藤克仁、
園尾聡、綱川隆司、中澤敏明、二宮崇、王向莉
「特許機械翻訳の課題解決に向けた機械翻訳技術解説」
自然言語処理 29巻 3号 pp. 925-985 (2022)
https://doi.org/10.5715/jnlp.29.925 （オープンアクセス）

発表の構成（≒論文の構成）

1. はじめに：解説論文の背景
2. 訳抜け・過剰訳への対策
3. 用語訳の統一
4. 長文対策
5. 低リソース言語対策
6. 評価
7. 翻訳の高速化・省メモリ化
8. おわりに：まとめ

1. はじめに
～本解説論文（技術調査）の背景

特許機械翻訳研究の動向

- 特許領域における機械翻訳の積極活用
  - 知財庁 (JPO, WIPO, USPTO, EPO, …)
  - 各種機械翻訳サービスと民間での活用
- 特許機械翻訳に関する学術研究
  - 2008-2013: NTCIR PATMT/PatentMT
    - 日本国内の統計的機械翻訳研究における中心的な役割を果たす
  - 2016-: WAT – JPO Patent Corpus
    - ニューラル機械翻訳への移行期に活用される
  - 以後は一般分野の対訳コーパスが充実し特許データの利用は減少
- 本解説論文では特許機械翻訳の諸課題の現状と今後を議論

2. 訳抜け・過剰訳への対策

訳抜け・過剰訳への様々な対策

- ネットワーク内で入力情報の参照度合いを管理
- 逆翻訳等で訳抜けを予測する仕組みの導入
- 語彙表現の工夫
  - 未知語や低頻度語が訳抜け・過剰訳の原因となることが多いため
- 出力長制御のモデル化
- 強化学習による訳抜け・過剰訳の抑制
- 繰り返しの抑制機構
- 訓練データのクリーニングとノイズ対策

3. 用語訳の統一

用語集や語彙制約の導入

- 対訳用語集に基づく後編集
  - 用語集定義の「避けるべき用語」を「仕様すべき用語」に置換
  - 同カテゴリの高頻度な用語に置換して翻訳し辞書情報で書き戻す
- 対訳辞書を活用するニューラル機械翻訳機構
  - 辞書中の訳語に対してスコア加点するモデル
  - 語彙制約つき探索
- 学習データへの語彙制約の適用
- 文脈を考慮した翻訳モデル
  - 前の数文の情報を参照する、過去の訳語をキャッシュする、…

4. 長文対策

長文に対する精度低下への対策

- 初期の改善：アテンション（注意）機構
  - 「全部覚えてから訳す」から「任意の入力時点を参照可能」に
- 文分割
  - 長い文を短い単位に分割して翻訳した後、結合する
    - ニューラル機械翻訳以降あまり進展がない
- 依存（係り受け）構造を制約にした翻訳モデル学習
- Transformerモデルの位置情報表現の工夫

※ 一般には学習時に長文が除外されることも多く注意が必要

5. 低リソース言語対策

（比較的）単純な方式

- カスケード翻訳：別の言語を経由した翻訳
  - 直接対訳の不足を経由言語（ピボット言語）のリソースで補う
    - 英語をピボットとすることが多いが、英→西→カタルーニャ語の研究も
- 学習データの拡張：機械翻訳によるデータの「水増し」
  - 目的言語の文章を原言語に翻訳（逆翻訳）
    - 目的言語側の表現は自然なので言語モデルの面では有効

最近の方式

- 転移学習
  - 別の言語対で作成されたモデルを所望の言語対で追加学習
    - 原言語、目的言語のいずれかを入れ替える
- 多言語モデル
  - 複数の原言語・目的言語の翻訳を単一モデルで行う
  - 副次的に低リソース言語の翻訳性能が向上する例も
- 事前訓練モデル
  - 大量のデータで学習することにより様々なタスクで有効
    - 単言語エンコーダモデル (BERT)
    - 多言語エンコーダ／デコーダモデル (mBART)

6. 評価

機械翻訳としての評価（人手評価）

- 段階評価
  - 忠実さと流暢さ、容認性
- 相対評価
  - 一対比較、ランキング
- 絶対評価
  - 直接評価（100点満点の点数評価）
- より系統的な評価
  - 多元品質評価尺度 (MQM)、JTF翻訳品質評価ガイドライン
  - 仕様に基づく体系、誤りの種別、深刻さによる重み付けスコア化

機械翻訳としての評価（自動評価）

- 表層ベース自動評価
  - BLEU、NIST、METEOR、TER、IMPACT、RIBES … 単語レベル
  - chrF、chrF++ … 文字レベル
- モデルベース自動評価（汎用モデルをそのまま応用）
  - BERTScore: BERTにより文類似度を計算
  - WE_WPI、MoverScore: 最適輸送を利用
- モデルベース自動評価（汎用モデルを評価にチューニング）
  - RUSE、BERT Regressor、BLEURT、C-SPEC、COMET

実務目線での評価

- NTCIR-9/10における特許審査評価 (2011, 2013)
  - 拒絶理由として用いられた既存発明の文書を機械翻訳し、拒絶判定根拠が得られるかを実務経験者が判定

- 特許庁「特許文献機械翻訳の品質評価手順」(2014)
  - 忠実さ・流暢さの段階評価 ＋ 重要技術用語の訳質評価
  - 誤りフィードバックのためのチェックリスト
  - 重要情報：技術的要素とそれらの相互関係
  - 国際ワークショップ WAT における JPO Adequacy として利用

7. 翻訳の高速化・省メモリ化

モデルの効率化

- モデルの小規模化
  - ニューラルネットワークのレイヤー数やベクトル次元数を減らす
- サブワード（単語より短い処理単位）
  - 単語間の共通箇所や活用語尾等を整理、語彙サイズ削減、効率化
- パラメータ共有
  - 同じ処理機構を再利用し、パラメータを増やさずに複雑な計算
- 自己アテンション（注意）機構の高速化 … 本来は長さの二乗
  - 注意可能範囲の制限、関数の簡略化
- 非自己回帰デコーダ
  - 先頭から順に予測、ではなく出力全体を並列予測

学習・実行の効率化

- 知識蒸留（teacher-student approachとも）
  - 小規模モデルに大規模モデルの出力を真似させるよう学習
- 探索空間の削減
  - ビームサーチのビーム幅を小さくし探索を簡略化・高速化
- 出力語彙の動的選択
  - 訳語予測時点の生成候補数を削減し計算量を削減
- ミニバッチ構成法の工夫
  - 同時に処理する文長を揃え、学習時の計算効率を最大化

実装による効率化

- 浮動小数点精度の削減：GPU向け
  - 単精度 (32bit) から 半精度 (16bit) に、計算効率も大きく向上
- 量子化：CPU向け
  - 行列乗算時に8ビット整数を利用
- CやC++による実装とメモリ管理
- 計算結果のキャッシュと再利用
- 翻訳コストを競う共通タスクも存在

8. おわりに

まとめと今後の展望

- 特許機械翻訳に関わる6つの課題に関する技術を解説
- 利用可能な技術は多く存在するが、実際に特許を対象に検証されている研究は少ない
  - 現在の多くの機械翻訳が共通ベンチマークデータに基づく
- 有望な要素技術は多く存在する
  - タスク間の知見共有と、特許機械翻訳への実用的展開が重要
  - 脱稿後も次々に新しい技術が発表
- 特許文書は特殊な外れ値ではなく、産業上重要な応用先

本活動について

- 研究会の知見集約と課題整理を目的に実施
  - 2019年12月　Japioとの議論により企画開始
  - 2020年4月　　各課題毎のサブグループに分かれ調査執筆開始
    （以後研究会定例会合にて進捗報告）
  - 2021年3月　　研究会年度報告書にて調査内容報告（中間）
  - 2022年3月　　研究会年度報告書にて調査内容報告（ほぼ最終）
  - 2022年5月　　「自然言語処理」解説論文として投稿
  - 2022年6月　　採録決定
  - 2022年9月　　出版（オープンアクセス）
  - 2023年3月　研究会年度報告書にて活動最終報告予定

謝辞と書誌情報（再掲）

- 論文共著者の皆様、当研究会関係者に御礼申し上げます
- 内容の詳細は論文をご参照ください

今村賢治、越前谷博、江原暉将、後藤功雄、須藤克仁、
園尾聡、綱川隆司、中澤敏明、二宮崇、王向莉
「特許機械翻訳の課題解決に向けた機械翻訳技術解説」
自然言語処理 29巻 3号 pp. 925-985 (2022)
https://doi.org/10.5715/jnlp.29.925 （オープンアクセス）