Webページ翻訳における単語単位の誤訳の修正

綱川 隆司　　梶 博行
静岡大学情報学部情報科学科

はじめに

- Webページの機械翻訳においては、原文の大意が把握できれば十分有用であるケースが多い
  - ー 必ずしも文法的または語用論的に高精度な翻訳結果が求められるわけではない
  - ー Webページの原文の大意を把握するのに重要な要素とは？
    - 内容語（特に名詞・動詞）の訳
    - 内容語間の関係を示す情報（主題、格情報、5W1H等）
    - レイアウト
    - 否定・同格（並列）など、文意を大きく変える文法的要素
    - …
- 原言語がある程度理解できるユーザの場合、内容語の訳を誤ると翻訳のメリットがなくなるということも・・・

- 多義性のある語subjectに着目し、Webページ翻訳結果の分析と、その修正可能性について検討する。

実験方法

1. Web検索(Google)を用いてsubjectをクエリとして検索し、上位1000件のページのうち以下のページを100件抽出する
   - subjectを含む30語以上の文章を本文として含む（科目リストなどページは除外）
   - 英語以外の言語を含まない
   - 辞書におけるsubjectという単語の説明ページではない
2. 該当ページにGoogle翻訳（英語→日本語）を適用し日本語ページを出力する
3. 該当ページの”subject”を含む文について統計的機械翻訳(Moses/Joshua)を用いて日本語文を出力する

Moses (Koehn et al., 2007), Joshua (Li et al., 2009) の訓練データとして日英新聞記事対応付けデータ (JENAAD)(Utiyama and Isahara, 2003)150000対訳を用いる

単語レベルの翻訳精度

- 抽出した100件のページのうち45件について、含まれるsubjectの訳語が全て正解、全て不正解、一部正解のいずれかで分類した
  - ー 1件のみ、2つの意味でsubjectが用いられているため、それぞれの正解訳で独立して扱った
  - ー 残りの44件については、subjectが複数の意味で出現しなかった (One sense per document (Gale et al., 1992))

subjectの正解訳	全部正解	一部正解	全部不正解	備考
件名	5	10	2	Email subjectが多数
対象	4	1	1	
主語	0	9	3	Subject-verb agreement等、手がかりとなる語が多い
主題	0	4	0	特許関係におけるsubject matter（主題）等
科目	0	3	4	大学等のページに多い
被験者	0	1	1	医学系、human subjects
テーマ	0	0	2	
被写体	0	0	1	写真の話題
サブジェクト	0	0	2	Subject directories、作品のタイトル
～について	0	0	1	On the subject of
～に従い、～を条件として	0	1	1	Be subject to（subjectは形容詞）、法律等の文書によく出現
合計	9	29	18	

訳語選択手がかりの分布

- 45件のページに含まれるsubjectについて、正解訳を導くために利用可能と考えられる手がかりをそれぞれ列挙した
  - ー 連語： 隣接する語と組み合わせて連語として扱うことで意味の特定が可能 （※下記の「文脈」も手がかりとなる場合はこちらに含む）
  - ー 文脈： 同一文書内に現れる語（同一文内とは限らない）から推定可能
  - ー 品詞： “be subject to”のsubjectは形容詞であり、形容詞という品詞から意味が推定可能

手がかり	全部正解	一部正解	全部不正解	備考
連語	4	16	4	Subject-verb agreement等
文脈	2	8	11	周辺に`sentence’等が現れる
品詞	0	1	1	Be subject to
その他	3	3	2	全部正解の例は「対象」が正解の場合
合計	9	29	18	

- 同一ページ内にsubjectが複数回出現した際、異なる訳語を割り当てる例が多かった
  - ー One sense per documentを適用することで、「一部正解」となっているページの多くは誤訳を修正可能
    - 複数の訳出のうち、どの訳語に統一するかは課題
    - より強力な手がかり（連語）から導かれた訳語に統一した場合、「連語」で一部正解になった16件のうち少なくとも12件は正解訳に統一可能であった
- 文脈を手がかりとして用いるには、語の翻訳確率モデルに（文単位を超えた）出現単語分布を導入する必要がある

Moses/Joshuaについて

- Mosesを用いて抽出されたフレーズ対
  - ー subject ||| ありうる, さ, され, され ず, される, と, とされ, とされる, とされるが, となる, なる, の 適用 外, ば, れ, れ ず, れる, 受け, 受ける, 問題, 対象, 対象 となっ, 対象 となる, 教科, 社会 が いかに 実効, 社会 問題, 科され, 適用 外
  - ー subject of ||| され ず, と
  - ー subject to ||| すると, なる, の 対象, の 対象 に なっ, れる, 受け, 受ける, 問題, 対象, 対象 と, 対象 となっ, 対象 となる, 対象 に なっ, 規制 の 対象 に
- Joshuaを用いて抽出されたSCFG対
  - ー [X,1] compulsory high school subjects ||| [X,1] 学力 低下
  - ー [X,1] high school subjects ||| [X,1] 低下
  - ー one subject ||| 見直し も
  - ー subject ||| 見直し
- Mosesの場合、大半が「対象」と訳されるかまたは正しく訳されなかった
- Joshuaの場合、一律「見直し」と訳された
- いずれの場合も訓練データの不足、および分野の不適合から生じていると考えられる。また、SCFG対の抽出は非常に困難であると予想される。

おわりに

- Webページ翻訳において大意の把握に重要な要素と考えられる内容語の翻訳を改善するため、多義語subjectに着目したWebページ翻訳結果の分析を行った
- 同一ページ内ではsubjectが単一の意味を持つ (One sense per document) ケースが大半だが、翻訳結果では同一ページ内で異なる訳を出力する場合が多かった
- Phrase-based統計的機械翻訳等を用いて連語の訳が既知であれば正しい訳が期待されるケースが多い。また、one sense per documentを適用すれば正しい訳に統一されることも期待される。
- 今後の課題
  - ー 文脈からの単語翻訳モデルの構築
  - ー 同一文書内での訳語統一手法の検討