10. 現状で利用可能なパラレル(対訳)コーパス

内山将夫@NICT
mutiyama@nict.go.jp

これまでのまとめ

コーパスベースの機械翻訳により
対訳コーパスから自動的に
翻訳機を作ることができることを述べ,
その手始めとして，IBM Model-1 を説明した.

ここでやること

ここでは，機械翻訳自体の話題は少し休んで,
コーパスベースの機械翻訳に必要な
対訳コーパスを
どう手に入れたら良いかを話す.

ここでの話題

- 対訳コーパスの重要性について
- 現状で利用可能な対訳コーパス
- 対訳コーパスを自動で作る方法
- 機械翻訳以外における対訳コーパスの利用

対訳コーパスとはなにか

複数言語について，特に，
意味内容がほぼ等しいと考えられる文について
対応関係が付いているコーパスを
パラレルコーパスあるいは対訳コーパスと呼ぶ

対訳コーパスの例

オオカミと仔ヒツジ
The Wolf and the Lamb
ある日のこと、オオカミは、群とはぐれて迷子になった仔ヒツジと出会った。
WOLF, meeting with a Lamb astray from the fold,
オオカミは、仔ヒツジを食ってやろうと思ったが、牙を剥いて襲いかかるばかりが能じゃない。
resolved not to lay violent hands on him,
何か上手い理由をでっち上げて手に入れてやろうと考えた。
but to find some plea to justify to the Lamb the Wolf's right to eat him.
そこで、オオカミはこんなことを言った。
He thus addressed him:
「昨年お前は、俺様にひどい悪口を言ったな！」
"Sirrah, last year you grossly insulted me."
仔ヒツジは、声を震わせて答えた。
「誓って真実を申しますが、私はその頃、まだ生まれていませんでした。」
"Indeed," bleated the Lamb in a mournful tone of voice, "I was not then born."
するとオオカミが言った。
Then said the Wolf,
「お前は、俺様の牧草を食べただろう！」
"You feed in my pasture."

対訳コーパスの重要性

- 対訳コーパスは，コーパスベースの機械翻訳にとって，前提条件である
- 歴史的には，
  - まず，対訳コーパスが自動構築され
  - つぎに，それを利用して，統計的機械翻訳が研究された

対訳コーパスはなぜ重要か

統計的機械翻訳では，入力文$\mathbf{f}$について

$$\hat{\mathbf{e}} = \arg\max_{e} P(\mathbf{e}|\mathbf{f})$$

なる$\hat{\mathbf{e}}$を出力するが，
この確率の推定には，手本となる対訳コーパスが必要だからである．

対訳コーパス構築の困難性

しかし，利用可能な対訳コーパスは少ない (後述)
その理由は，対訳コーパスの構築には，様々なことを解決しないといけないからである．

対訳コーパス構築における障害

- 原文および翻訳文の獲得
  - 原文および翻訳文には，著作権保持者がいる．
  - この著作権保持者の許可を得ないと
  - その文章は対訳コーパスに採用できない
  - たくさんの著作権保持者と交渉するのは，時間も費用もかかる
- 高価

  たとえ対訳コーパスがあったとしても

  それらは，しばしば，高価である．

研究利用可能なコーパスの例

- Linguistic Data Consortium のコーパス
  年会費2500ドル(30万弱)を払うことにより,
  - 中国語 - 英語
  - アラビア語 - 英語
  - フランス語 - 英語

  の対訳コーパスを入手可能である．各言語対について100万文以上ある．
- Europarl コーパス
  欧州諸語についてのコーパスが無償で利用可能である．各言語対について数十万の規模である．
- NICTで公開しているコーパス (無償)
  日本語と英語の対訳コーパス．30万文程度である．
- NTCIRで利用可能なコーパス (無償)
  NICTで開発した日米特許対訳コーパス180万文を利用して，特許翻訳タスクが実施されている．

NICTで無償公開している対訳コーパス

- 読売新聞とThe Daily Yomiuriの対訳文18万文
  http://www2.nict.go.jp/x/x161/members/mutiyama/jea/index-ja.html
- 小説など160作品の対訳約10万文
  http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html

日英新聞記事対応付けデータ

「読売新聞」と「The Daily Yomiuri」とで互いに翻訳関係にあるような文を自動的に対応付けたデータ

- 1対1文対応：15万，1対n文対応：3万
- 日英2言語コーパスとしては世界最初のある程度の規模のコーパス
- 研究教育目的に利用可能

文対応の精度は 98% 程度

サンプル

- 欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。 Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen.
- 我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。We recognize the importance of improved market access for economic progress in Russia.
- 法人レベルでのパートナーシップ及びマネージメント支援は、特に効果的であり得る。Partnerships and management assistance at corporate level can be particularly effective.

検索例

low profile

「Japan has kept a low profile」「日本は目立った行動を控えていた」

dispose of

「dispose of bad assets」「不良資産の処理」

データの特徴

- 新聞記事
- 高品質な実例

欲しい表現全てを網羅するほど大きくはないが，頻出表現は網羅できる．
→ コーパスに基づく英語教育に利用

日英対訳文対応付けデータ

再配布可能な日英の作品について，対訳文対応を1文単位で付けたデータ

- Project Gutenbergや青空文庫やプロジェクト杉田玄白などの作品について
- 160作品を公開

Project Gutenberg

Project Gutenberg, abbreviated as PG, is a volunteer effort to digitize, archive and distribute cultural works. Founded in 1971 by Michael Hart, it is the oldest digital library.[1] Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of August 2007, Project Gutenberg claimed over 22,000 items in its collection. Project Gutenberg is affiliated with many projects that are independent organizations which share the same ideals, and have been given permission to use the Project Gutenberg trademark.

From Wikipedia, the free encyclopedia

青空文庫

青空文庫は，利用に対価を求めない，インターネット電子図書館です．著作権の消滅した作品と，「自由に読んでもらってかまわない」とされたものを，テキストとXHTML (一部は HTML) 形式でそろえいています．
「青空文庫早わかり」より

プロジェクト杉田玄白

プロジェクト杉田玄白というのは、いろんな文章を勝手に翻訳して公開しちゃうプロジェクトなのだ。プロジェクトグーテンベルグや、青空文庫の翻訳版だと思って欲しい。日本は翻訳文化だといわれるけれど、それならいろんな翻訳が手軽に入手できるようにすることで、もっともっと文化的な発展ができるようになるだろう。

作品の一部

「80日間世界一周」「DESのクラック：暗号研究と盗聴政策、チップ設計の秘密」「RMS スウェーデン王立工科大学講演」「「年」の話」「『群集心理』」「あらし」「ひと，場所，もの，そして，アイデア」「わがままな大男」「アッシャー家の崩壊」「アモンティリャードの酒樽」「アラビー」「アルセーヌ・ルパンの逮捕」「イソップ寓話集」「エヴリン　「ダブリンの人々」より」「オズの魔法使い」「カール・マルクス Interview」「カウンターパーツ」「キリストにならいて」「クリスマス・カロル」「グレイト・ギャツビー」「グロリア・スコット号」「ケンジントン公園のピーターパン」

データの特徴

- 小説やエッセイ

楽しみのために読むことができる.

デモ