1. オープンソースのMTシステムの例

内山将夫@NICT
mutiyama@nict.go.jp

機械翻訳手法の概要

- 人手による規則に基づく機械翻訳 (商用の機械翻訳システムのほぼ全て)
  日：主語 目的語 述語 → 英：主語 述語 目的語
  私は本を読んだ → I read a book
- コーパスに基づく機械翻訳 (現在の研究の主流)
  用例：私は本を読んだ → I read a book
  入力：私は新聞を読んだ → 出力：I read a newspaper.

コーパスベースの機械翻訳の主構成要素

- 対訳コーパス：対訳コーパスがあれば，翻訳エンジンを利用して，機械翻訳ができる．しかし，これがなければどうしようもない．→ コーパスベースの機械翻訳のボトルネック.
- 翻訳エンジン：対訳コーパスから単語や句の対訳を抽出し，それを利用して翻訳をする．現状の研究の主対象.
- 翻訳精度の評価手法：翻訳エンジンの精度を自動評価できれば，その評価を最大化するように翻訳エンジンを改良できる.

オープンソースの機械翻訳システム

http://www.statmt.org/wmt07/baseline.html

- 対訳単語等の抽出ソフトウェア
- 機械翻訳エンジン (当該ページに対訳コーパスへのポインタ)
- 翻訳精度の自動評価とそれを利用したパラメタ調整

このページだけで一応の機械翻訳ができる (学生実験レベル)
機械翻訳研究におけるベースライン.

ベースラインシステムの性能

- 解析速度：一文あたり数秒
- 翻訳精度：語順が似た言語間については，市販のシステムと同程度？
- 翻訳可能な言語：3000 万単語 (100 万文) 程度の対訳コーパスがある言語対

機械翻訳エンジンの現状と今後の課題

現状

大量の対訳コーパスが必要

対訳コーパスと異なるドメインの翻訳は苦手

主に構造が似た言語対の翻訳で実験されている

課題

少量の対訳コーパスからの学習

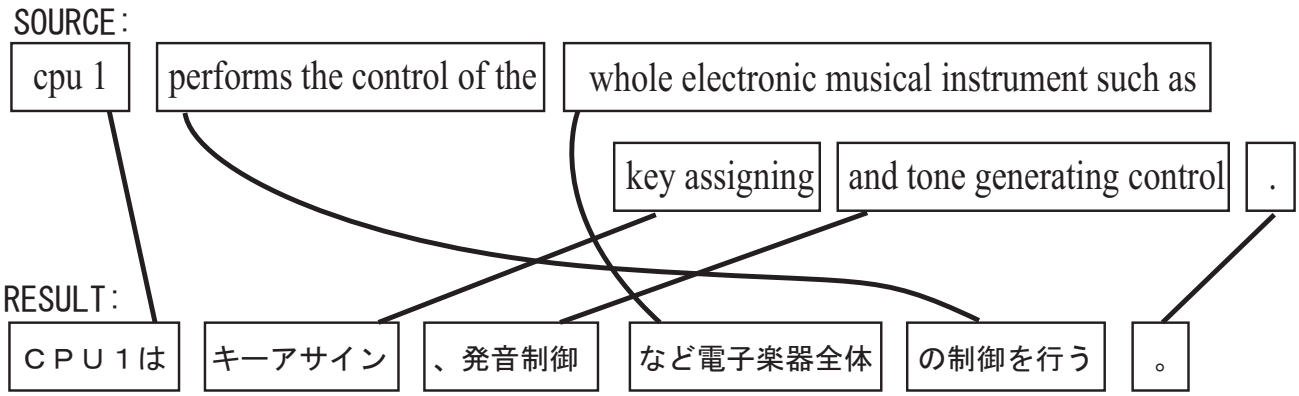
異なる分野への翻訳エンジンの適応

異なる構造の言語での翻訳エンジンの評価

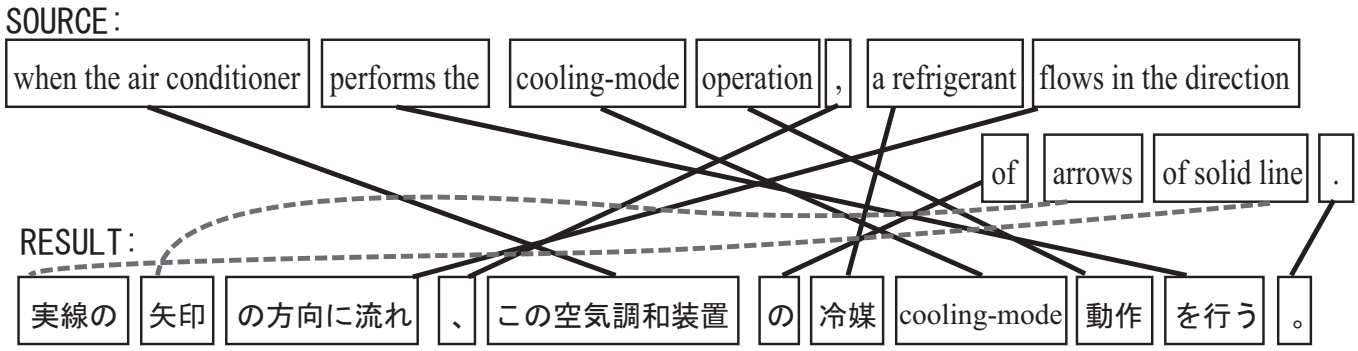
日本語のコーパスの例

コーパスベースの機械翻訳には，数十万から数百万の対訳文が現状では必要である．これに匹敵する大きさの対訳コーパスは日英では，日本と米国に同時出願された特許から構成されたものしかない．
これは，NICTが開発したものであり，700万文程度の日英の対訳文からなり，
国立情報学研究所(NII)主催のNTCIRプロジェクト
http://research.nii.ac.jp/ntcir/index-ja.html
の特許の機械翻訳タスクにおいて利用されている．
2008年12月以降に，一般にも，NIIから研究利用可能となる予定であるが，現在のところは，特許の機械翻訳タスクの参加者のみが利用可能である．

翻訳例 (1/4)



(a) 長いフレーズがヒットして正しい翻訳に成功している例（↑）



(b) フレーズはおおよそ正しいが、並び替えで失敗している例（↑）

翻訳例 (2/4)

SOURCE: consequently , the potential of internal data transmitting line io becomes higher than the potential of internal data transmitting line / io .
REFERENCE: この結果、内部データ伝達線ＩＯの電位が内部データ伝達線／ＩＯの電位よりも高くなる。
RESULT: これにより、内部データ伝達線ＩＯの電位が内部データ伝達線／ＩＯの電位よりも高くなる。

翻訳例 (3/4)

SOURCE: it is necessary to select the first and second magnetic layers 1 and 2 in order to obtain a large gmr effect .
REFERENCE: 第１及び第２磁性層１，２は高いＧＭＲ効果のために選ばれる必要がある。
RESULT: 第１及び第２の磁性層１と２を得るために、大きなＧＭＲ効果を選択する必要がある。

翻訳例 (4/4)

SOURCE: when the air conditioner performs the cooling-mode operation , a refrigerant flows in the direction of arrows of solid line .
REFERENCE: 空気調和機が冷房運転する場合、冷媒は実線矢印の向きに流れる。
RESULT: 実線の矢印の方向に流れ、この空気調和装置の冷媒 cooling-mode 動作を行う。

まとめ

ある程度の性能の計算機さえあれば，誰でも機械翻訳の研究ができるようになった.

課題

興味があれば,
http://www.statmt.org/wmt07/baseline.html
のシステムを動かしてみる．(これはそれなりに大変です)
参考情報としては,
NTCIR-7特許翻訳タスクのページ
http://if-lab.slis.tsukuba.ac.jp/fujii/ntc7patmt/index-ja.html
があります.