3. MTの性能をどう測定するか？ 実験の作法とMTの自動評価

内山将夫@NICT
mutiyama@nict.go.jp

実験における性能評価

MTの性能をどう測定するか？

- 実験の作法
- MTの性能の自動評価

実験の目的

MTの性能を正確に測定したい

Q: 正確とはどういうことか？

A: この実験で得られた結論が，別の実験でも成立することが，だいたい言えること．

つまり，結論が，一般化できること．

実験の作法

- 実験データを，訓練用，パラメタ調整用，テスト用に3分割する
- 訓練データで，MTモデルの基本構造を得る.
- パラメタ調整用データで，モデルのパラメタを調整する
- テストデータで，モデルの性能を確認する

モデルの性能を正しく測定するには

- 訓練，調整，テストで，データに重なりがあってはいけない．なぜなら，たとえば，
- 訓練とテストに同一のデータがあったら，単に，
- そのデータを記憶しておくだけで，高性能となる．しかし，
- 単に記憶するだけでは，未知データに対処できない．つまり，
- 一般性が低い．それにもかかわらず，
- 高性能と判断されたら，
- そのテストはおかしい．

モデルの性能を正しく測るには

色々なデータを使う必要がある
MTの場合には以下のようなものを翻訳したい

- 多言語：日英，日中，日韓,...
- 多ジャンル：新聞，論文，特許，ブログ，チャット,...

現在の技術では，ジャンルが変わっただけで，性能が大きく低下する場合が多い.
その理由：

- ジャンルが変わると，出現する単語や句が変わる

実際にはどのような実験がなされているか

- 訓練，調整，テストは必ず分ける
- できるだけたくさんの種類のデータを使う
  → しかし，
  - 実験に利用できるデータは少ないし，
  - 実験には，時間と手間が掛かるので，
  - それらを勘案して，
  - できるだけのことをする．

ともかく実験をしたとしてMTの性能をどう測るか？

2つの軸がある

- 人手評価か，自動評価か
- MT訳自体の評価か，MT訳により何ができるかによる評価か

MT訳自体の評価が評価に利用される場合が多い

考えられる理由

- MT訳自体の品質があがれば，MT訳を使ってできることの効率も上がると考えられるから
- 研究開発においては，タスクにかかわらず，MT訳を向上させることができれば，それの方が手間がかからないから

しかし，MTにより何ができるかによる評価を，もっと盛んにする必要がある．なぜなら，それこそがMTを使う理由だからである．

課題3

複数のWeb上の MT システムについて,「こういう用途なら，システムAよりもシステムBの方が良い，なぜなら,... だから」というように評価してみる．

MT訳自体の評価の方法

- 人手評価
- 自動評価

人手評価の例

- 翻訳文の流暢さ (読み易さ)
- 翻訳文の忠実度
  どのくらい原文に沿っているか？

などを各5点満点で評価するなど.

自動評価の例

- 参照用の翻訳と MT 訳との類似度を
- n-gram (n 単語連鎖) の
- 重なり具合で評価する.

参照訳と似ている訳が良い訳であるという立場である.

人手評価の良いところ

- 明確な指示を評価者に与えなくても，何らかの良さの評価ができる
- 自動評価ではできない細かな評価ができる
  文の類似性の判定には，n-gram の一致だけでなく，類似語とか，言い換えとかも考えないといけない.

人手評価における研究課題

ある評価者がある文の読み易さを中程度などと判定したとき，なにゆえ，そのような判定をしたかを調べること．たとえば，

- 間違った助詞を選択したとか
- 時制が違うとか
- 係り受けが違うとか
- 語順が違うとか

様々な要因が読み易さにはあるが，どのような要因があるかは，まだリストアップされていない．

人手評価の問題点

時間がかかる

- 1000 文の MT 訳について，その訳が良いか悪いかを判定するには
- 一週間以上かかるかもしれない.

一方,

- MT システムを開発するときには,
- すぐさま，評価結果が欲しい

自動評価ができれば，この問題は解決する.

自動評価の良いところ

すばやくできる

モデルのパラメタを調整するときに，調整用データにおけるMT訳の品質が向上する．つまり，自動評価の値が大きくなるように，パラメタを調整できる．

評価の安定性

同じMT訳と参照訳について，同一の値がでる．
→ 異なるシステムの比較が容易である．
人手評価だと，同じ人が同じ文を評価しても，異なる結果になることがある．

自動評価の悪いところ

- 自動評価では，測定できないものがある．
  たとえば，BLEU という評価尺度では，ngram の重なりしかみていないので，同一の意味でも異なる表現の場合には，間違いとされる．
- しかし，自動評価は
  - 似たタイプのシステムの比較には有用である．

まとめ

- 訓練，調整，テストにデータを分け，
  - 訓練で，モデルの基本を作り
  - 調整で，パラメタを調整し
  - テストで，テストする
- 調整にあたっては，自動評価の値が最大となるように，パラメタを調整する
- 自動評価には，限界もあるが有効なツールである．