2. MTの性能評価についての一般的な話題

内山将夫@NICT
mutiyama@nict.go.jp

コーパスベースのMTをするときに問題になること

- MTの性能をどう測定するか？
- MTの計算モデルをどう捉えるか？
- MTの計算モデルをどう訓練するか？
- MTの訓練に必要なコーパスをどう獲得するか？

これらについて，この講義では，実例を示す.

MTの性能をどうはかるか？ — その前に —

Q: 何故 MT の性能を測定する必要があるか？

A: MTシステム A と B とを比較するとき，A と B のどちらが良いかを知るためには，A と B の性能を測定する必要がある．

Q: システムの比較はなぜ必要か？

A: より良いシステムを作るためには，システムを比較し，どちらが良いかを知る必要がある．

MTの性能は相対評価で測る

相対評価とは2つのシステムをある基準により比較すること

Q: ある基準とは何か？

A: MTを使う目的により異なる．よく使われる基準については後述する．

MTにおける絶対評価とは何か？

- よくわからないが,
- ある目的について，そのMTが
- 使えるか？使えないか？の評価だろう.
- たとえば，ケータイ翻訳で，そのケータイ翻訳だけで
- 海外旅行ができるかどうかとかが
- 絶対評価で,
- 相対評価では，2つのシステムの両方が使えなくても,
- まだ，こちらの方がましということがある.

何故相対評価をするか？

- たとえ，2つのシステムの双方が使えないにしても
- その2つのシステムを比較することにより，
- どちらのシステムの方が良いかが分かれば，
- その良い方のシステムを残して，
- それを更に改良することができるため

システムを少しずつ改良するためには，相対評価で十分である．

評価の基準は目的による

- たとえば，英文を書くときに，
- Web 上の MT システムを利用するとしたら，
- その英文作成に最適なものが良い．
- この基準は，難しいが，
- MT 訳の，ある種の「良さ」を評価するものである．

なるべく目的独立の基準が欲しい

- MTを使う目的には様々なものがあるので
- それぞれの目的ごとにMTを評価する必要があるのは
- その目的のためにMTを使う人にとっては当然である.
- しかし，MTを開発するときには,
- 全ての目的を考慮するのは無理なので,
- ある基準をもってきて，その基準を満せば,
- いろいろな目的が果せるようなものがあると良い.

課題1

- 目的独立の基準の欠点
- 目的依存の基準には，どのようなものあるか？

MT訳と参照訳が似ていれば良いのではないか?

- MTは入力文を翻訳する.
- この翻訳文が正しいものであれば，MTを使って,
- 各々の目的を達成できると思われる.
- したがって，MT訳と正しい訳(参照訳)を比較して
- それが似ているほど良いと考える.

MT訳と参照訳が似ているとはどういうことか?

- たとえば，参照訳とMT訳とを見比べて，
- 誤訳になっている個所を指摘したりする．
- あるいは，共通する単語数を数えたりする．
  → 後述する自動評価

事例：各種 Web MTシステムの比較

- 新聞記事からとった10文について,
- Q:入力英語
- A:参照訳
- として，S1,S2,S3のシステムを比較する
- 比較には，誤訳の数を数えることにする.

Q1: Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen.

A: 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。

S1: ヨーロッパは活発にエディンバラで同意されて、コペンハーゲンで強化されたGrowth Initiativeを実行しています。

S2: ヨーロッパは、活発に、エジンバラで同意されて、コペンハーゲンで強化されるGrowth Initiativeを実行しています。

S3: ヨーロッパは、エディンバラで同意され、コペンハーゲンで強くなった成長イニシアチブを活発に実行しています。

誤訳の数

S1: 3.「活発に」の位置.「Growth」と「Initiative」が未訳.

S2: 4.「活発に」の位置.「強化される」が誤訳.「Growth」と「Initiative」が未訳.

S3: 0.

Q2: We recognize the importance of improved market access for economic progress in Russia.

A: 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。

S1: 私たちはロシアに経済進歩のための立直り市況アクセスの重要性を認めます。

S2: 我々は、ロシアにおける経済進歩のために、改善された市場参入の重要性を認めます。

S3: 私たちは、ロシアで経済進歩のために改善された市場参入の重要性を認識します。

誤訳の数

S1: 3.「ロシアに」と「立直り」と「市況」が誤訳

S2: 0.

S3: 1. 「ロシアで」が誤訳

Q3: Partnerships and management assistance at corporate level can be particularly effective.

A: 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。

S1: 法人のレベルにおけるパートナーシップと管理支援は特に有効である場合があります。

S2: 会社レベルの協力と管理援助は、特に効果的でありえます。

S3: 企業のレベルの協力および管理援助は特に有効になりえます。

誤訳の数

S1: 0

S2: 0

S3: 0

Q4: The Federal Constitutional Court decides on the question of unconstitutionality.

A: 違憲の問題については、連邦憲法裁判所が決定する。

S1: 連邦政府の Constitutional Court は違憲の問題を決めます。

S2: Federal Constitutional 法廷は、憲法違反の問題を決定します。

S3: 連邦憲法裁判所は、憲法違反の質問を決めます。

誤訳の数

S1: 3. 「Constitutional」と「Court」が未訳.「問題を」が誤訳.

S2: 3. 「Federal」と「Constitutional」が未訳.「問題を」が誤訳.

S3: 1. 「質問を」が誤訳.

Q5: The practical process of integration must begin in the economic sphere.

A: 統合の実際のプロセスは、経済分野から始めねばならない。

S1: 統合の実用的な過程は経済球で始まらなければなりません。

S2: 統合の実用的なプロセスは、経済球で始まらなければなりません。

S3: 統合の実際的なプロセスは経済球体の中で始まるに違いありません。

誤訳の数

S1: 2. 「実用的な」と「球」が誤訳

S2: 2. 「実用的な」と「球」が誤訳

S3: 2. 「球体」と「違いありません」が誤訳

Q6: Sanctions should be upheld until the conditions in the relevant Security Council resolutions are met.

A: 関連する安全保障理事会決議の諸条件が満たされるまで、制裁は維持されるべきである。

S1: 関連安全保障理事会の決議における条件が満たされるまで、制裁は是認されるべきです。

S2: 関連した安全保障理事会決議の状況が対処されるまで、制裁は支えられなければなりません。

S3: 適切な安全保障理事会の決議中の条件が満たされるまで、制裁が支持されるべきです。

誤訳の数

S1: 2. 「関連安全保障理事会」と「是認」が誤訳

S2: 3. 「状況」と「対処」と「支えられ」が誤訳

S3: 1. 「適切な」が誤訳

Q7: International terrorism is a grave threat to world peace and security.

A: 国際テロは、世界の平和と安全に対する重大な脅威だ。

S1: 国際テロは世界の平和とセキュリティへの危険な脅威です。

S2: 国際テロは、世界平和と安全に対する重大な脅威です。

S3: 国際テロは世界平和とセキュリティに対する重大な脅威です。

誤訳の数

S1: 2. 「セキュリティ」と「危険な」が誤訳

S2: 0.

S3: 1. 「セキュリティ」が誤訳

Q8: Poverty, population policy, education, health, the role of women and the well-being of children merit special attention.

A: 貧困、人口政策、教育、保健、女性の役割、及び児童の福祉は、特別の注意に値する。

S1: 貧困、人口政策、教育、健康、女性の役割、および子供の幸福は特別な注意に値します。

S2: 貧困、人口方針、教育、健康、女性の役割と子供たちの幸福は、特別な注意に値します。

S3: 欠乏、人口政策、教育、健康、女性の役割および子供の安寧は、特別の注意に値します。

誤訳の数

S1: 0.

S2: 1. 「人口方針」が誤訳

S3: 1. 「欠乏」が誤訳

Q9: Improvement of access for Russian products to international markets strongly reinforces Russian structural reform.

A: 国際市場に対するロシア産品のアクセス改善は、ロシアの構造改革を大いに強化する。

S1: 国際市場へのロシアの製品のためのアクセスの改良は強くロシアの構造改革を補強します。

S2: 国際的な市場へのロシアの製品のためのアクセスの改善は、強くロシアの構造改革を補強します。

S3: 国際市場へのロシアの製品のためのアクセスの改良は強くロシアの構造の改革を強化します。

誤訳の数

S1: 0

S2: 0

S3: 0

Q10: This also means respecting power structures established in a democratic way.

A: このことはまた民主的な形で樹立された権力構造の尊重をも意味する。

S1: また、これは、民主的な方法で確立された権力機構を尊敬するのを意味します。

S2: これも、民主主義の方向で設立される権力側を尊重することを意味します。

S3: これはさらに民主主義の方法で設立された権力機構を尊敬することを意味します。

誤訳の数

S1: 0

S2: 4. 「これも」と「方向」と「設立される」と「権力側」が誤訳

S3: 0

誤訳の集計

- S1: $3+3+0+3+2+2+2+0+0+0=15$
- S2: $4+0+0+3+2+3+0+1+0+4=17$
- S3: $0+1+0+1+2+1+1+1+0+0=7$

文単位の比較

S3 $>$ S1 $>$ S2 という傾向がありそうだ.

S1 の方が S2 より良い文の数 4 S2 の方が S1 より良い文の数 2	S1 $>$ S2
S1 の方が S3 より良い文の数 1 S3 の方が S1 より良い文の数 5	S3 $>$ S1
S2 の方が S3 より良い文の数 2 S3 の方が S2 より良い文の数 4	S3 $>$ S2

今の比較の問題点

- テスト文が少ない．10文では，十分な比較はできない．
- テスト文が恣意的である．新聞記事からとった文では，新聞記事以外の翻訳については，評価が不十分である．
- 「誤訳」の定義があいまい．なんとなく誤訳では，客観的な評価と言えない．

これらの問題点の逆を考えると

- 十分な数の文数が必要
- 恣意的でないテスト文が必要
- 「誤訳」の定義を客観的に確立する．

これらが成立してはじめて，システムの公正な比較ができる．しかし，これを全てするのは困難である．

評価に関しての最近の傾向

１つのテストセットに対して，複数の研究機関が参画する共同タスクが盛んとなっている．

- 同一データにより，異なる手法を比較できる．
- 共同のデータを利用して，よりよい評価を研究する．

良いテスト文があったとして，どれくらいの差があれば良いか？

- 比較の方法として，10文について，各文ごとに
- システムＡとＢを比べたとき，Ａが，たとえば，
- 6文について，Ｂより良かったとする．すると，
- Ａ対Ｂ＝6対4である．このとき，
- Ａの方がＢよりも良いシステムであると言ってよいか？

良くない

- なぜなら，10回中6回くらいは,
- 偶然かもしれないからである．一方,
- 10回中8回なら，これは,
- 偶然の可能性は低い．このようなことを測るために,
- 統計的検定を利用する.

Q: なぜ差があるかを比較するか？

- A: もし差があれば,
- システムを良い方向に改良するし,
- 差がなければ,
- その変更は，受け付けない.
- 差があるかどうかを知ることにより,
- システムを，その方向に変更すべきかどうかが分かる.

符号検定

- 簡単な検定方法として，符号検定がある．
- これは，システムＡとＢが同じ性能だとすると，
- ある文について，Ａが良い確率は0.5であることに基づく．
- 0.5の確率のとき，10回中6回Ａが良い確率は，

$$_{10}C_6 0.5^{10} = 0.20508$$

- である．そして，0,1,2,3,4,5回のそれぞれについては
  0.00097656，0.00097656，0.043945，0.11719，0.20508，0.24609
  である．
- したがって，Ａが0〜6回，Ｂより良い確率は，

$$\sum_{i=0}^{6} {}_{10}C_i 0.5^{10} = 0.82812$$

  である．一方
- Ａが 7〜10回良い確率は，
  1 - 0.82812 = 0.17188
  である．

確率の判断の仕方

- 6回よりもAが良い確率が
  $0.17188\ (= 17\%)$
  とかなりあるので，AとBに性能差があるとは言えない.
- 一方，8回のときには，Aが9,10回良い確率は
  $$1 - \sum_{i=0}^{8} {}_{10}C_i 0.5^{10} = 0.0107(1\%)$$
  なので,
- 10回中8回Aが良い場合には,
- そうなる確率が小さいので，統計的に有意差があると言う.

統計的検定における注意点

- AとBの性能に差があるということだけを言っていて,
- その差は，とても小さいかもしれない.
- たとえば，10000回中で，5100回，Aが良かったら,
- Aが良い割合は，0.51で,
- AとBとの差は小さいが,

$$1 - \sum_{i=0}^{5100} {}_{10000}C_i 0.5^{10000} = 0.022(2.2\%)$$

なので，有意差はある.

- つまり，有意差があるということと
- その差の大きさが十分なものかは
- 別問題である.
- また，テスト文自体が良いものかのチェックも必要である.

まとめ

- ここまでで，システムの性能を
- 相対評価で比較すること
- 評価の際には，テスト文の選び方が大切なこと
- システム間の差が有意かどうかを
- 統計的に検定できることがわかった.

課題2

- Web 上の MT システムを複数選び，その性能を比較すること