テストセットを用いた日中翻訳エンジン評価

2012年9月7日

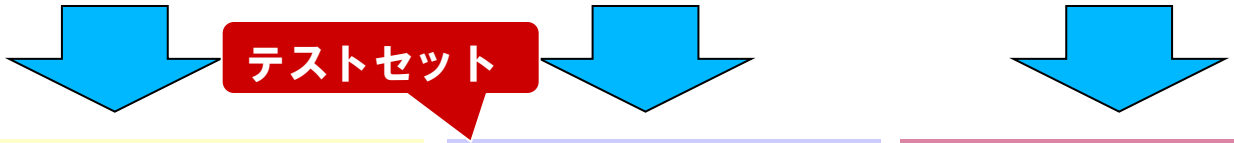
㈱富士通研究所
長 瀬 友 樹

評価は誰のために　何のために

研究者　・方式優位性確認
・コンテスト型WS

開発者　・目標設定　・開発への
・性能改善の確認　フィードバック
・他社比較

利用者　・導入システム決定の参考　・コスト削減の裏づけ



相対評価
（従来型自動評価、人間評価）

エラー分析
（項目別評価）

実業務評価
（目的別・業種別、チューニング評価）

テストセットとは

文法項目ごとに設問（チェックポイント）が付いた基本例文集
原文の機械翻訳結果に対して、評価者がYes/Noで評価

[例]

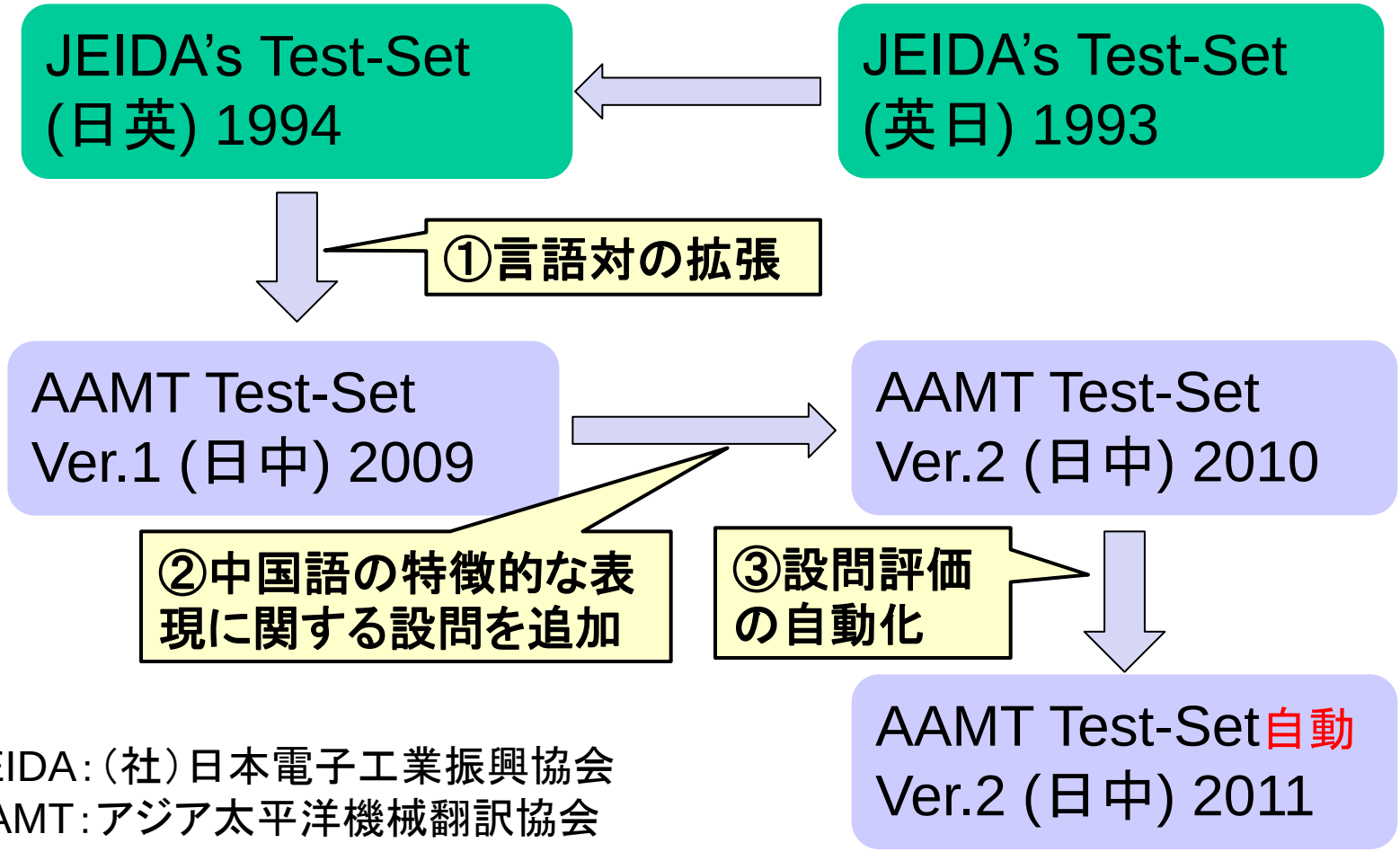
文法項目	原文	訳文（参照訳）	設問（チェックポイント）
比較	彼は君より高い	他**比**你高	比較文に「比」を使っていますか
比較	これはあれより大きい	这个**比**那个大	比較文に「比」を使っていますか
比較	これはあれより大きくない	这个**没有**那个大	比較文の否定は「没有」になっていますか
比較	彼女と同じくらい綺麗だ	**跟**她一样漂亮	「同じくらい」が「跟…一样」になっていますか

本発表では、テストセットを用いる評価 ＝ 設問ベースの評価

テストセットを用いて
- 翻訳システムごとに項目別の得点を算出　⇒エラー分析
- 翻訳システム全体の訳文品質を定量化　⇒相対評価

AAMT日中翻訳テストセットの開発と検証実験

AAMTテストセットの開発プロセス



JEIDA:(社)日本電子工業振興協会
AAMT:アジア太平洋機械翻訳協会

①言語対の拡張（日中翻訳対応）

■ 英語訳を中国語訳に置換え設問を見直し

原文のカテゴリ　原文　正解（参照訳）　チェックポイント（設問文）

カテゴリー		日本文	中文１	設問（日本語）	設問（中国語）
（1）述部	1	彼は多くの研究者を集めた。	他使很多的研究者聚集起来。	「集めた」の部分の自動詞/他動詞用法の訳し分けは正しいですか？	“集めた”部分的自动词/他动词的译法是否正确？
（1－1）述部の訳し分け	2	彼は標本を集めている。	他在收集标本。	自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
	3	彼は論文を集めて本にした。	他把论文收集成册。	自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
	4	彼らは会議室に集まった。	他们在会议室集合。	自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
	5	学生が教室に集められた。	学生在教室里集合。	自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法及被动句的翻译是否正确？
（1－2）断定文	6	この装置はバッテリ駆動だ。	这个装置是电池驱动的。	判断文の訳文は正確ですか？	判断句的翻译是否正确？
	7	手順は左右同一である。	程序是左右相同的。	判断文の訳文は正確ですか？	判断句的翻译是否正确？

②中国語の特徴的表現への対応

- 39項目、251文の設問を追加

中国語の特徴的な表現の評価が可能に

1「の」と「的」	14 許可（可以）	27 補語
2「Ｖた」と「的」	15 願望	28 使役文
3「もの」と「的」	16 前置詞	29 受身文
4方位词	17 比較文	30 被を用いない受身文
5疑問詞	18「Ｖたばかり」と「刚刚」	31 可能
6否定文（否定副詞の位置）	19「～ている」と「了」「着」	32 自発
7「不」と「没」	20「了」（変化）	33 尊敬
8「在」文	21「～ている」が無標になる場合	34 敬語
9「有」文	22「過」（～たことがある）	35 ようだ・そうだ
10 形容詞述語文	23 自他動詞	36 「う・よう」と「吧」
11 兼語文	24 ～で	37 「～ば～ほど」と「越～越～」
12 二重目的語をとる動詞	25 のだ文	38 介詞
13 できる・可能（「会・能・可以」）	26 「把」字文	39 決まり文句

テストセット中の例文数（設問数） 325（Ver.1）⇒576（Ver.2）

③設問評価の自動化

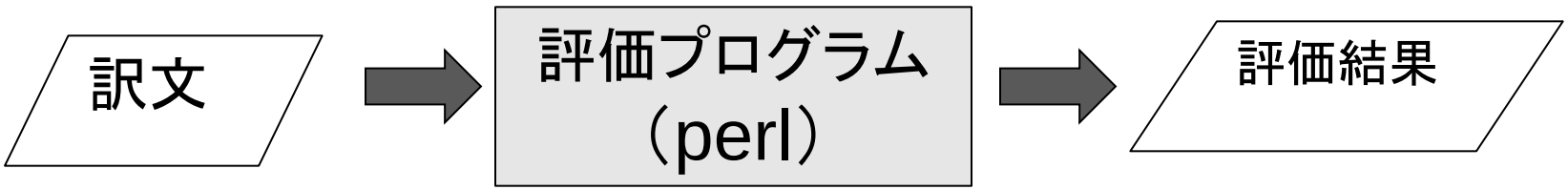
- テストセットの各設問を正規表現へ書き換え

例文：　冷たいものをください
正解：　我要凉的
設問：「冷たいもの」が「凉的」になってますか

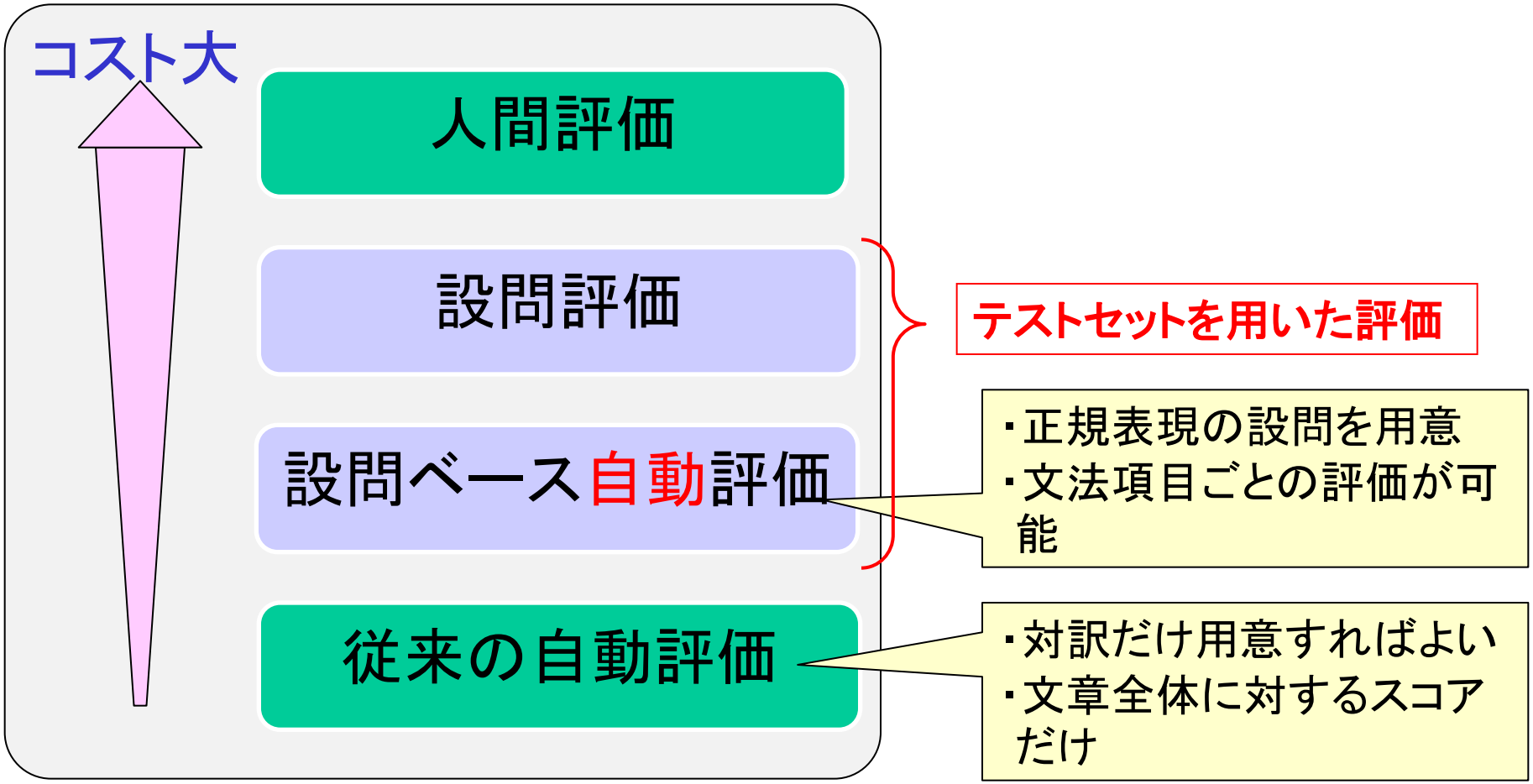
```
elsif($_ =~ /我要.+的$|给我.+的$/){
    if($_ =~ /凉的|冷的|冰的|冰镇的/){
        print $_, "¥tYes¥n"
    }
...
```

書き換えのときに同義語候補を追加

- 正規表現をもとに設問評価を自動評価できるプログラムを作成



設問ベース評価の位置づけ



設問評価の検証実験

(1) 人間評価:

正確さ評価
流暢さ評価
5段階(1–5)評価

設問ベース評価
Yes/no評価

(2) 自動評価 :

既存の自動評価
(BLEU, NIST, WER, PER)

設問ベース評価
(自動評価)

※ test set : 251 例文 (39 文法項目)

System A
System B
System C
System D
System E
System F

6種類の
代表的日中システム

人間評価と自動評価の相関

評価方法	正確さ	流暢さ
BLEU	0.9453	0.9588
NIST	0.9783	0.9123
1-WER	0.9801	0.9267
1-PER	0.9707	0.8986
設問ﾍﾞｰｽ評価（人間評価）	0.9793	0.9334
設問ﾍﾞｰｽ評価（自動評価）	0.9902	0.9208

- どの組み合わせでも高い相関を示している
- 設問ベース評価は人間評価、自動評価ともに、従来の評価指標と比べて、信頼性に遜色はない

設問の回答パターン分布

自動-人間（評価者1）

一致	Yes-Yes	425	1181	78.4%
	No-No	756		
相違	Yes-No	222	325	21.6%
	No-Yes	103		

自動-人間（評価者2）

一致	Yes-Yes	473	1179	78.3%
	No-No	706		
相違	Yes-No	174	327	21.7%
	No-Yes	153		

人間（評価者1）-人間（評価者2）

一致	Yes-Yes	456	1264	83.9%
	No-No	808		
相違	Yes-No	170	242	16.1%
	No-Yes	72		

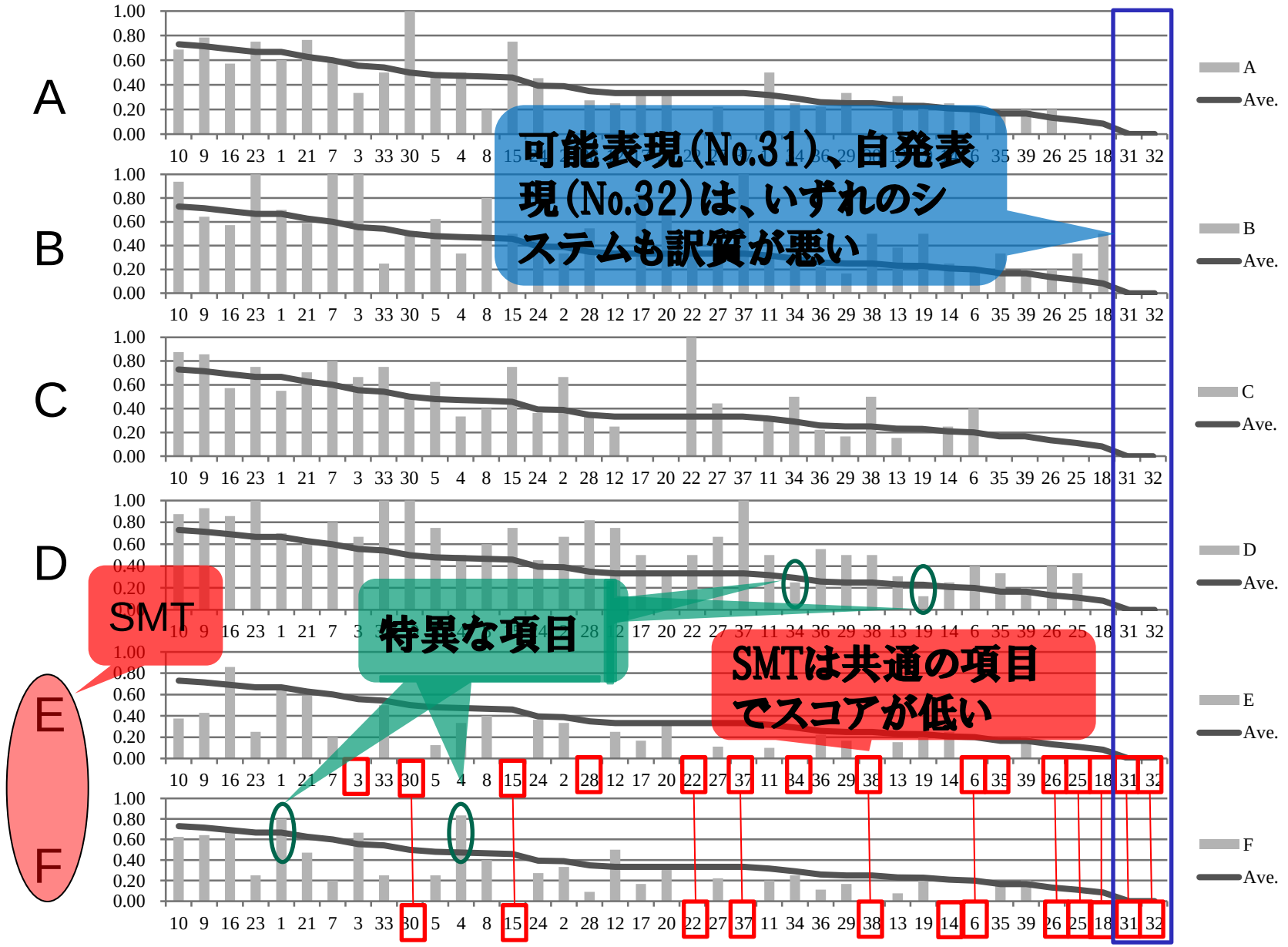
設問の人間評価と自動評価の一致度（Kappa係数）

比 較 対 象	Kappa係数
自動評価　: 人間評価（評価者1）	0.5494
自動評価　: 人間評価（評価者2）	0.5552
人間評価（評価者1）: 人間評価（評価者2）	0.6616

- 設問ベース自動評価は、人間評価と一致している（有意水準1%で検定）

  ⇒　設問ベース自動評価を人間評価の代わりに用いることが可能

システム毎の文法項目別スコア



まとめ～テストセット評価のメリット

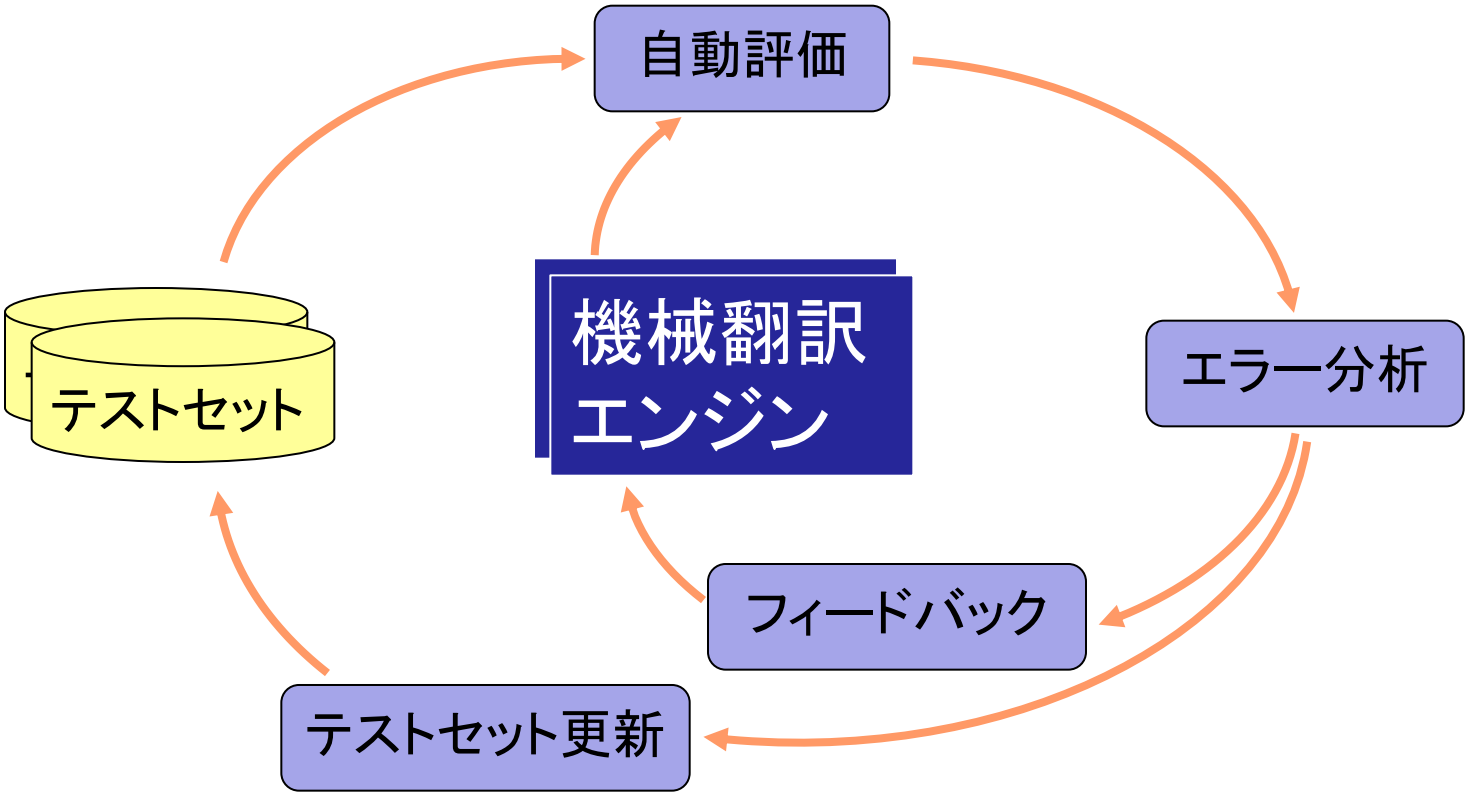
- システムごとに文法項目別のフィードバックが得られる（結果を見てエラー分析ができる）
  - 特に開発ベンダにとっては注力すべきポイントを決める上で重要な情報が得られる
- 評価の自動化が可能
  - 設問を正規表現に近似して自動化した評価結果は、人間による設問評価の結果と有意に一致する
- 従来の人間主観評価と高い相関
  - 他の自動翻訳と比べても遜色のない性能が期待できる

まとめ～テストセット評価の課題

- 言語対に特有の文法項目への対応
  - 解析側と生成側の双方の言語に特有の文法事項に対応する必要がある
- システムの成長に合わせた設問の改変
  - 開発当初は基本文中心、性能向上に合わせて特殊な文法項目に関する設問に移行
- 開発工数の効率化
  - テストセット開発には文法項目の選択、例文と設問の設定など工数がかかる
- テストセットへのチューニングを回避できる運用
  - チューニングしてしまうと同じテストセットは使えない

テストセットのライフサイクル
（開発者向けの理想的評価環境）

テストセットのライフサイクル



評価結果をエンジン開発にフィードバックできることがテストセット評価の特徴
このサイクルを効率的に回すことで機械翻訳エンジンの開発が加速する
　今後の研究課題　：　テストセット作成（更新）の自動化
　　分析ツールの可視化　etc.

[参考]
日中／日英機械翻訳エンジンの比較

[参考]日中/日英エンジンの比較（1）

	日英翻訳評価（377文評価の平均）		
	Adequacy	Fluency	設問
評価者１	2.94	3.20	0.75
評価者２	3.59	4.25	0.68
平均	**3.27**	**3.73**	**0.72**

	日中翻訳評価（377文評価の平均）		
	Adequacy	Fluency	設問
評価者3	2.87	3.36	0.31
評価者4	2.39	2.41	0.48
平均	**2.63**	**2.89**	**0.40**

※Adequacy/Fluencyは評価点として1-5の整数を付与　　設問はYes=1,No=0として集計

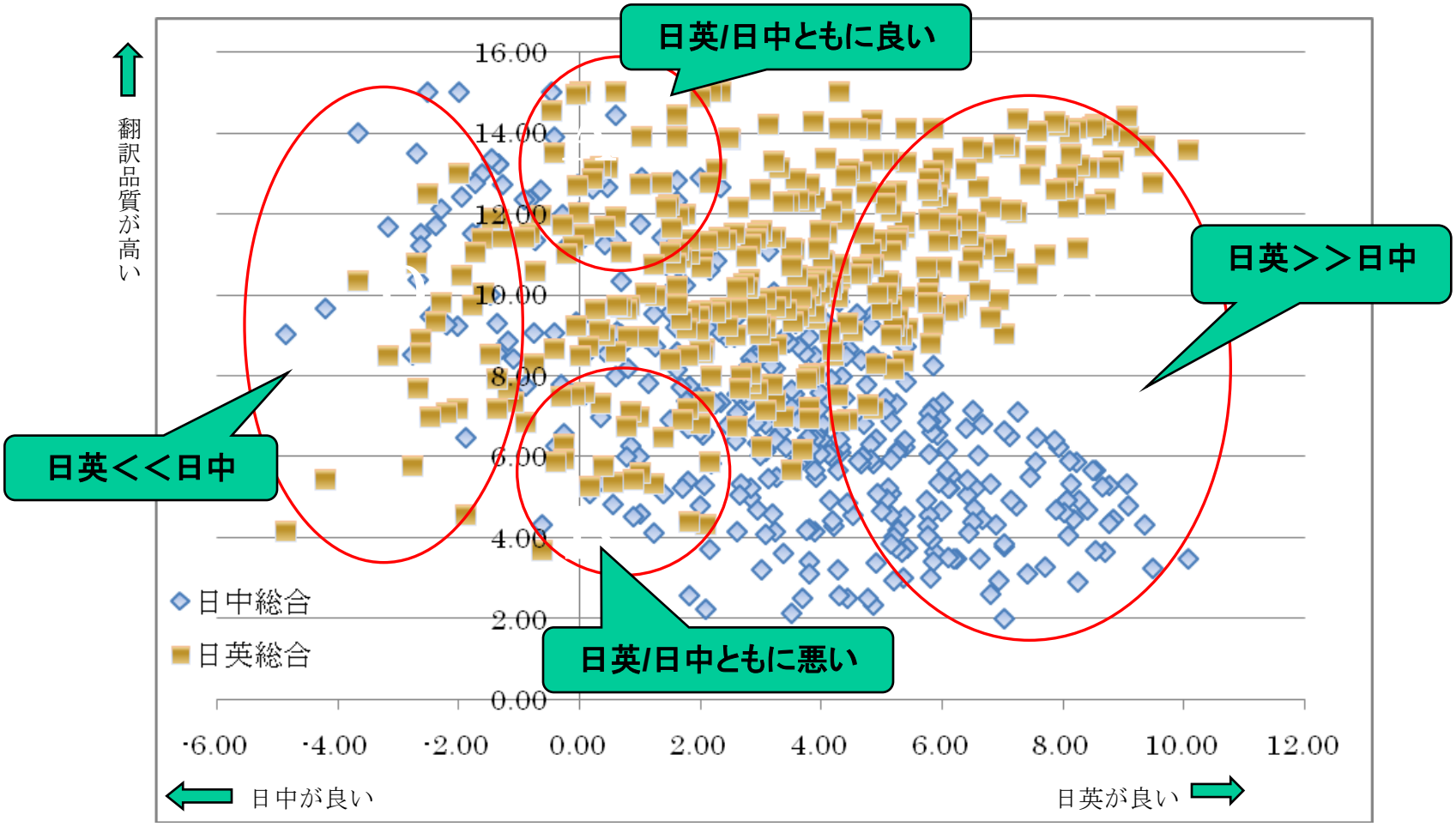
	日英/日中
Adequacy	1.24倍
Fluency	1.29倍
設問	1.80倍

Adequacy、Fluencyの5段階評価は、評価者が無意識のうちに評価の平均が真ん中に近づくことがあるかもしれない。

⇒設問評価の方が、システムの優劣がよりはっきり表れる?

[参考]日中/日英エンジンの比較（2）

全評価文について、Adequacy,Fluency,設問（5点換算）の合計値（縦軸）と日英/日中の合計値の差分（横軸）の位置にプロット　⇒圧倒的に日英の評価値が高い



[参考]日中/日英エンジンの比較（3）

日中　＞＞　日英　（Dゾーン）

数は少ないが日英翻訳特有の課題を含む文がこのゾーンに入ると思われる。

日中翻訳では問題とならないが、日英翻訳で訳し分けが必用になる文が存在する。

例：「渋滞が自然解消する。」

日英翻訳で「自然」を”by itself”と訳すのは、なかなか難しい。
日中翻訳では、そのまま「自然解消」でOK。

⇒テストセットを用いて対象言語対の特徴分析も可能