

評価値・機械翻訳修正訳付き 日英中韓対訳コーパスの構築

藤田 篤 隅田 英一郎

情報通信研究機構

今日の話題

■ MT訳の評価と修正訳を含む対訳コーパスの構築

- ねらい: MT品質の推定 & MTユーザの満足度向上
- エントリの例

日本語原文	片道だけで買えますか。
模範訳	May I get it for one way?
MT訳 (new)	Can I buy just one way?
評価値 (new)	B
修正訳 (new)	Can I just buy a one way ticket ?

- 日本語 → 英語, 中国語, 韓国語
 - 日本語原文: 旅行会話8789発話, 病院内会話1,676発話
 - 句に基づくSMT (PBSMT)でMT訳を生成
 - 人手で模範訳, 評価値, 修正訳を付与

背景と動機

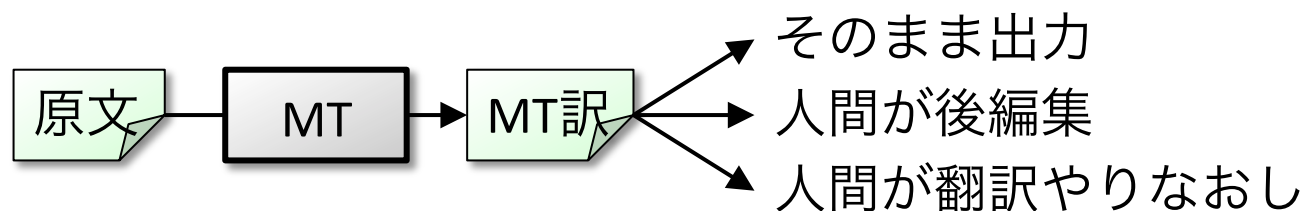
MTを実社会で活用する際の課題

■ MT訳の品質のばらつきへの対処

- MTによる高品質な翻訳は不可能 [Bar-Hillel, 51]
- 人間が、適切に使いこなす必要がある
 - 人間翻訳の道具として [Hutchins+, 92]
 - 一般のユーザによる即時利用も要考慮

■ MT訳の品質推定(Quality Estimation; QE) [Specia+, 10]

- 複数のシステムの出力からの選択 [He+, 10]
- 後編集の負荷の見積もり [Specia+, 11]
- ユーザに対して品質を自己申告 [Specia+, 11]
 - e.g., 自動結果の直接使用の可否 [Soricut+, 10]



MT訳の品質推定

- MT時の尤度(自信)との違い
 - ユーザの立場での品質
- BLEU等の自動評価尺度との違い
 - 参照訳を用いない
 - 原文を参照する
- 問題設定 [Bojar+, 16]
 - 語レベル: 個々の語の適否
 - 句レベル: 問題のある範囲の同定
 - 文レベル
 - 想定されるMT訳後編集の割合 (HTER) [Snover+, 06]
 - 正確さや流暢さに基づく絶対評価
 - 文書レベル: 要旨を掴む目的の使用可否

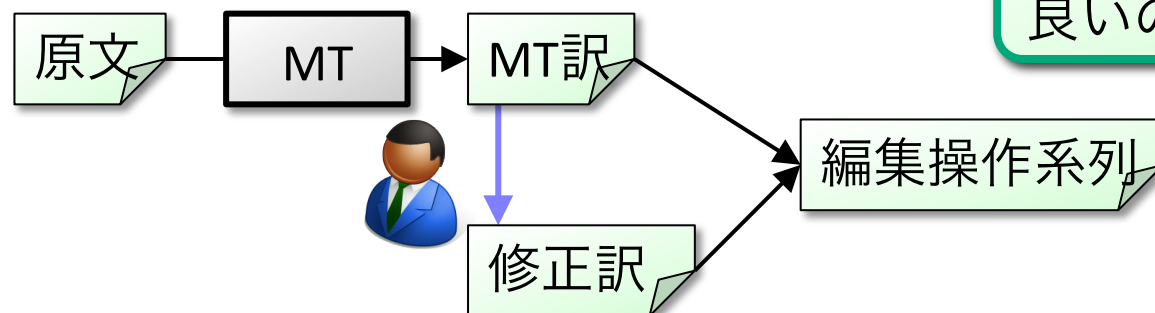
既存の訓練・開発・評価用データ

■ 欧州言語の一部に限定

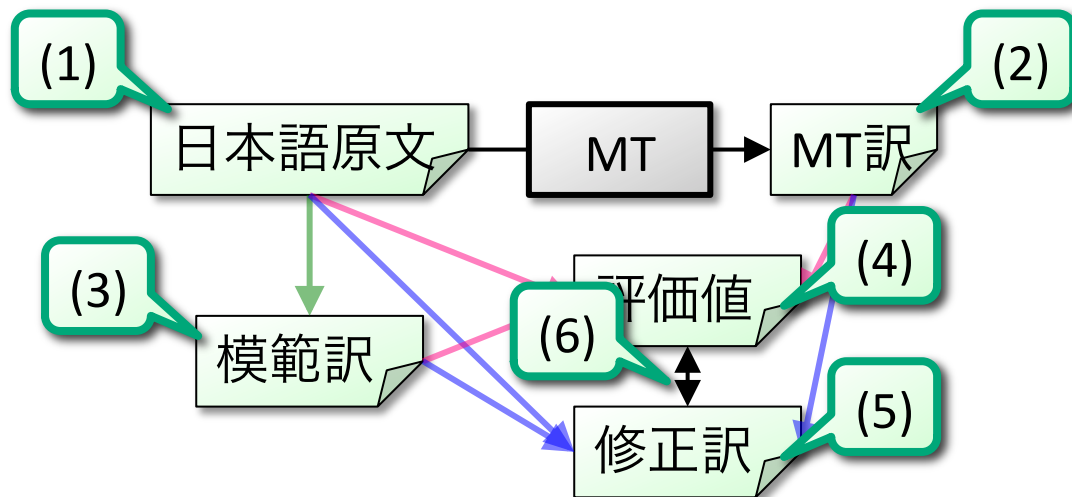
出典	言語方向	分野	文数		
			train	dev	test
[Potet+, 12]	Fr→En	ニュース	10,881	0	1,500
[Bojar+, 15]	En→Es	ニュース	11,271	1,000	1,817
[Bojar+, 16]	En→De	IT	12,000	1,000	2,000

■ MT訳の後編集を想定することが主流

- 語レベルQE: 後編集結果の修正訳から自動タグ付け
- 文レベルQE: 上記結果に基づく HTER



コーパスの構築手順 (6ステップ)



(1)と(2)



■ (1) 日本語原文の取得

- 2つの分野の話し言葉
 - 旅行会話: VoiceTraの実入力の自動認識結果
 - 病院内会話: 音声翻訳の模擬被験者実験における実発話
- 人手でフィルタリング
 - 意味が分からない発話
 - 公序良俗に反する発話
 - 音切れや咳払いなど

■ (2) MT訳の生成

- 句に基づくSMT (PBSMT)
 - 日→英: VoiceTra 201310 (701k文対で訓練)
 - 日→中, 日→韓: VoiceTra 201609 (1.63M文対で訓練)

(3) 模範訳の作成

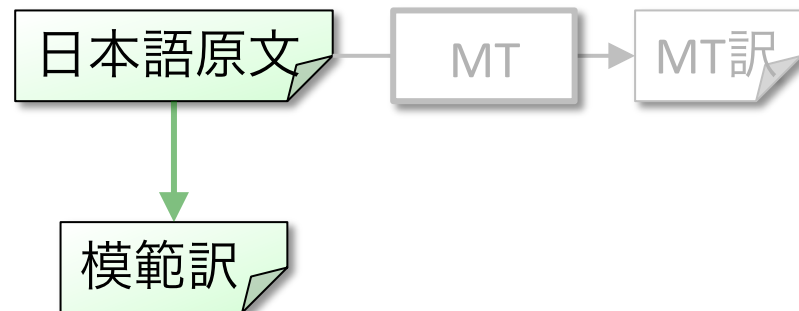
■ 参照情報: 原文のみ

■ 発話ごとに文脈独立に翻訳

- 必要に応じて発話の状況を想定
- 明らかではない情報を付加してはならない
- 省略されている代名詞等を補う場合は統一的に

■ チェッカーを立てて納品前検査 (予稿の図2)

- 内容
 - 過不足のなさ
 - スペルチェック
 - 句読点や文末記号の適切さ
- 形式・スタイル
 - 引用記号, 全角/半角, スペースなどの統一的使用
 - 見えない文字(末尾のスペース, C2A0, 改行コード)の削除



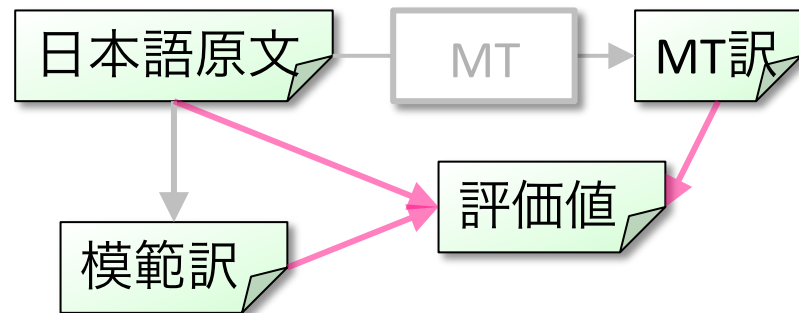
(4) MT訳の評価値の付与

■ 参照情報

- 日本語原文, 模範訳, MT訳

■ 6段階評価 (予稿の表2)

- S: ネイティブ並み
 - A: 申し分ない
 - B: まずまず
 - C: 許容範囲
 - D: 意味不明
 - E: 訳文なし
- 修正不要
- 要修正



評価値の種類と判断基準 (修正不要)

■ S (ネイティブ並み)

- 原文の情報が漏れなく翻訳されており，訳文に文法的な間違いがない．使われている語彙もネイティブから見て自然である．

■ A (申し分ない)

- 使われている語彙はネイティブから見て不自然であるが，原文の情報が漏れなく翻訳されており，訳文に文法的な間違いがない．句読点(あるいはピリオド，コンマ)の抜けやアルファベットの大文字と小文字の違い等の軽微な誤りも，意味が通じるならば許容する．

評価値の種類と判断基準 (要修正)

■ B (まずまず)

- 原文のあまり重要でない情報が一部漏れていたり，訳文に綴り誤りや，文法的な誤り，句読点の誤りがあったりするが，容易に理解できる。

■ C (許容範囲)

- 原文の重要な情報が漏れていたり，訳文に文法的な間違いが多数あったりしてかなり崩れた訳文ではあるが，原文の意味をなんとか理解することはできる。

■ D (意味不明)

- 重要な情報が誤訳されており，原文の意味を正しく理解することは不可能である。

(5) MT訳の修正

■ 参照情報

- 日本語原文, 模範訳, MT訳

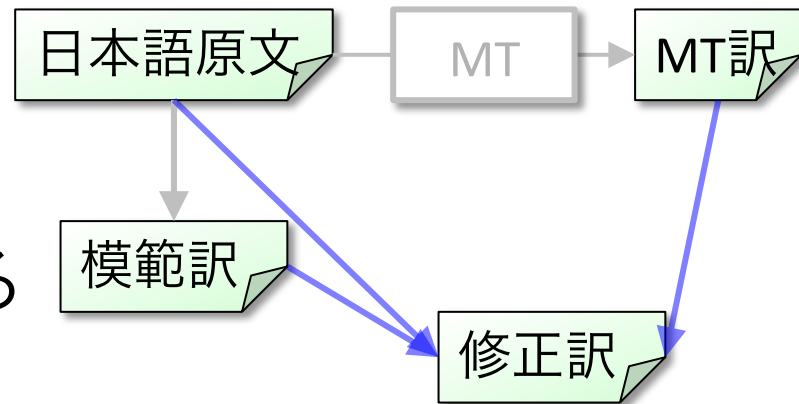
■ 必要最低限の修正でS/Aにする

- 基本操作単位

- 語の削除: 1語削除 = コスト1
- 語の挿入: 1語挿入 = コスト1
- 語の置換: 1語置換 = コスト1
- 語句の移動: ひとまとまりの語句の移動 = コスト1

- 編集回数(MT訳 → 修正訳) ≤ 編集回数(MT訳 → 模範訳)

- 模範訳よりもMT訳から遠くない
- “<”だと制約としては厳しすぎる
- 模範訳をコピーして修正訳とするのはNG



編集回数は自動カウント

■ TERのツール: tercom-0.7.25

- 分かち書き
 - 英: Mosesのトークナイザ
 - 中: 文字
 - 韓: MeCab + mecab-ko-dic (平文中のスペースは特別視)
- 修正訳作成時: 編集回数(MT訳 → 修正訳)
- HTER = 編集回数(修正訳 → MT訳) / (修正訳の語数)
 - 語句の移動を考慮すると、向きによって回数が異なりうる

入力文	あなたは22時までにはチェックインする必要があります。									
MT訳	You should check in by twenty two o'clock .									
修正訳	You have to check in by 22:00 .									
	=	Sub	Del	=	=	Sub	Ins	Ins	=	

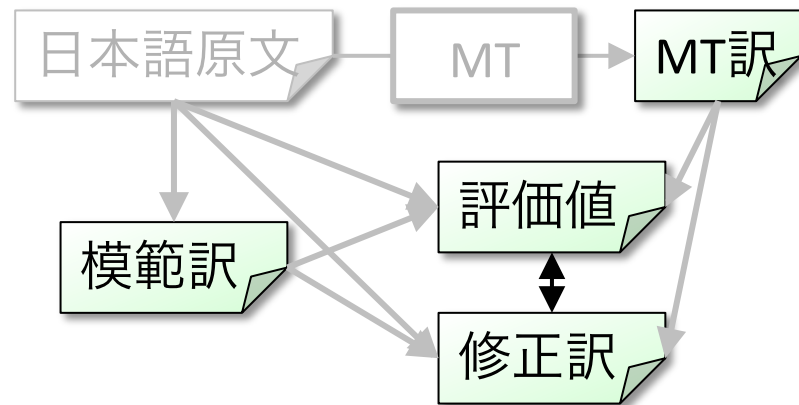
(6) 整合性の確認 (自動)

■ 参照情報

- 評価値, 修正訳

■ 下記のいずれかに該当する場合は(4)と(5)をやり直し

- 修正の要不要が一貫していない
 - 評価値S/A, MT訳!=修正訳
 - 評価値B/C/D, MT訳==修正訳
- 修正が必要最低限でない
 - 編集回数(MT訳 → 修正訳) > 編集回数(MT訳 → 模範訳)



構築したコーパスの諸元

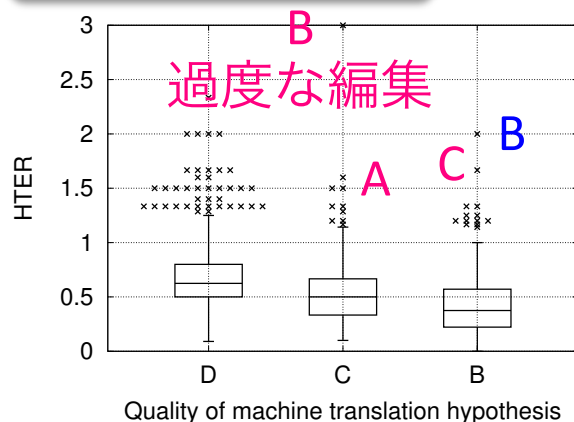
日→英中韓のコーパス

- 独立に構築した日英・日中・日韓を突合
- 評価値ごとの文の数と割合 (予稿の表3)
 - 日英(古いモデル): 全体的に性能がいまいち
 - 日中・日韓(新しいモデル): 結構高精度

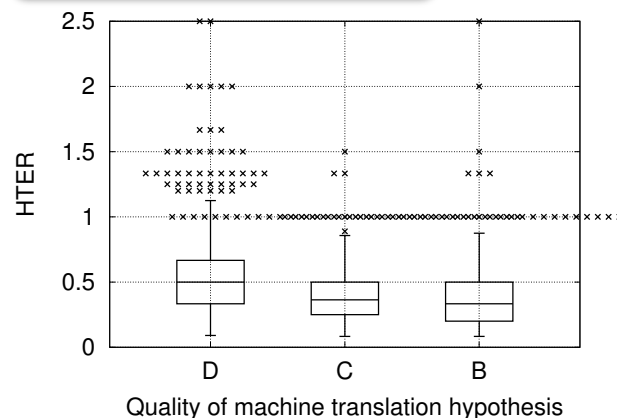
	旅行会話 (8789発話)						病院内会話 (1676発話)					
	日英		日中		日韓		日英		日中		日韓	
評価値	文数	割合	文数	割合	文数	割合	文数	割合	文数	割合	文数	割合
S	1961	22.3%	2829	32.2%	3466	39.4%	95	5.7%	708	42.2%	903	53.9%
A	1462	16.6%	1875	21.3%	2326	26.5%	107	6.4%	514	30.7%	482	28.8%
B	1269	14.4%	1275	14.5%	1361	15.5%	181	10.8%	172	10.3%	166	9.9%
C	1067	12.1%	899	10.2%	724	8.2%	333	19.9%	107	6.4%	97	5.8%
D	3026	34.4%	1909	21.7%	908	10.3%	960	57.3%	175	10.4%	28	1.7%
訳なし	4	0.0%	2	0.0%	4	0.0%	0	0.0%	0	0.0%	0	0.0%

評価値とHTERの関係

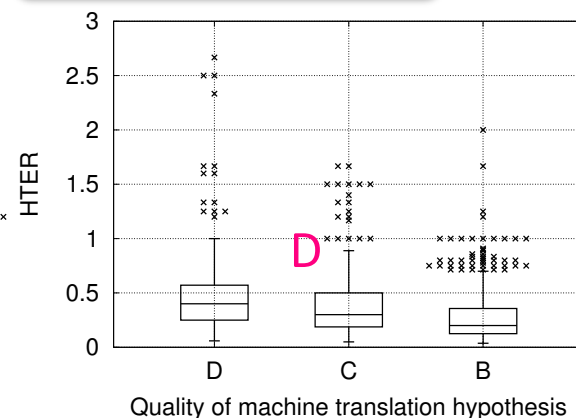
日英, 旅行会話



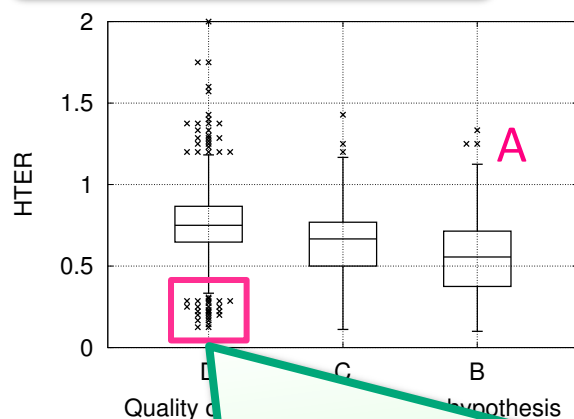
日中, 旅行会話



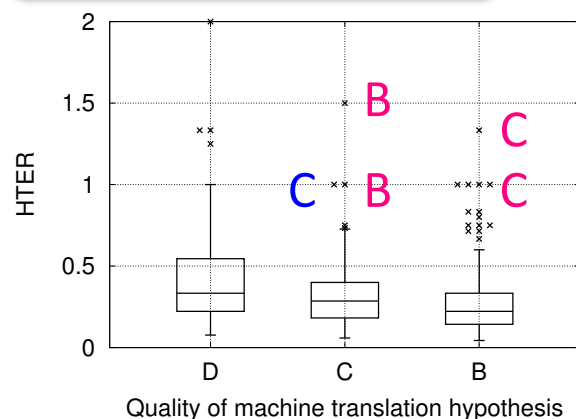
日韓, 旅行会話



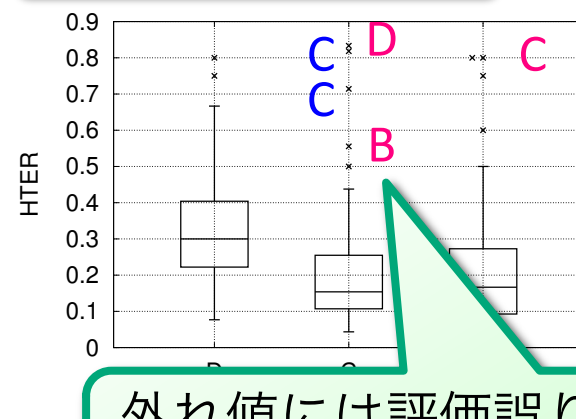
日英, 病院内会話



日中, 病院内会話



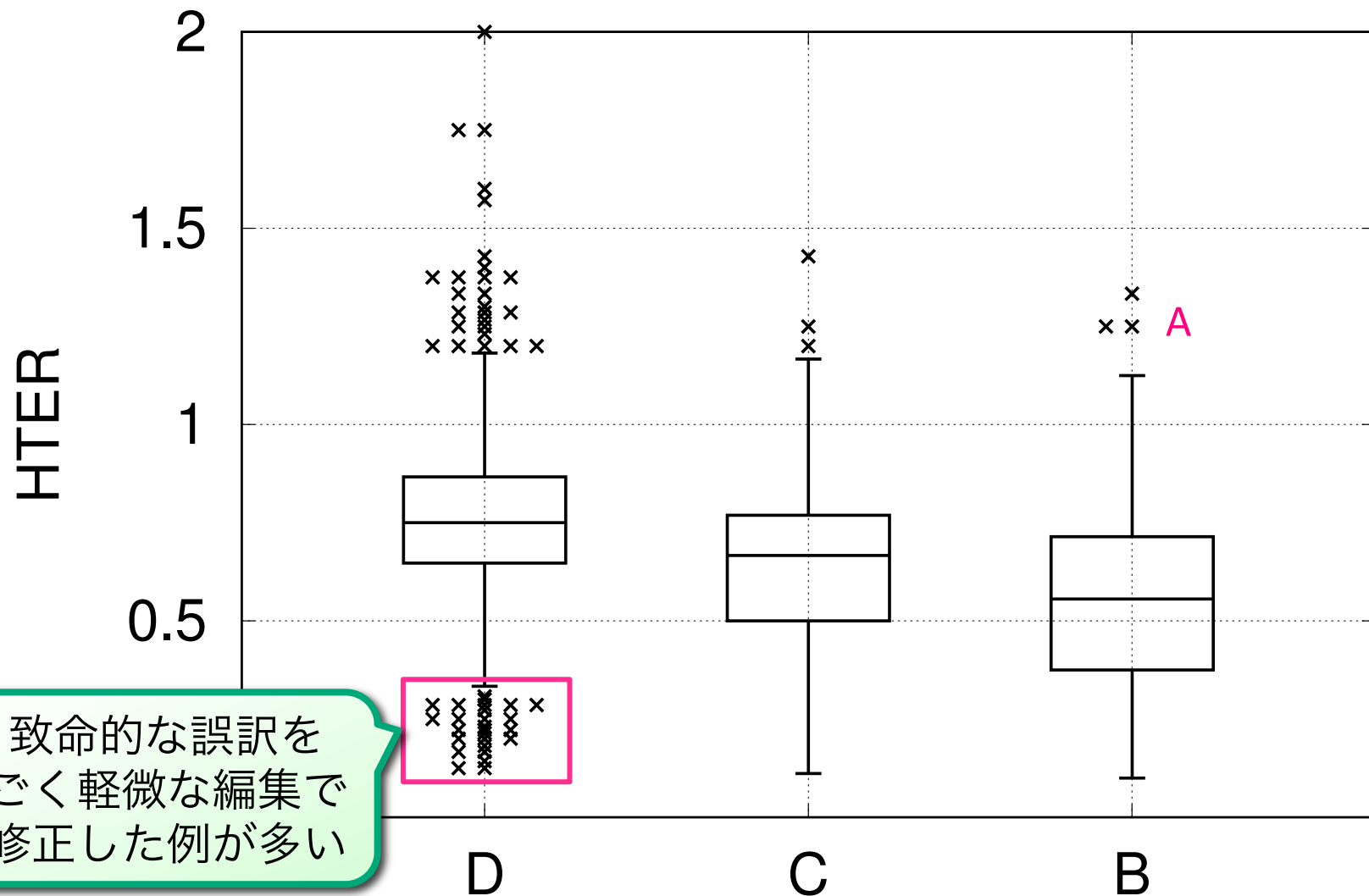
日韓, 病院内会話



致命的な誤訳をごく軽微な編集で修正した例が多い

外れ値には評価誤りが多い

評価値とHTERの関係 (日英, 病院内会話)



評価値がDでHTERが低い例

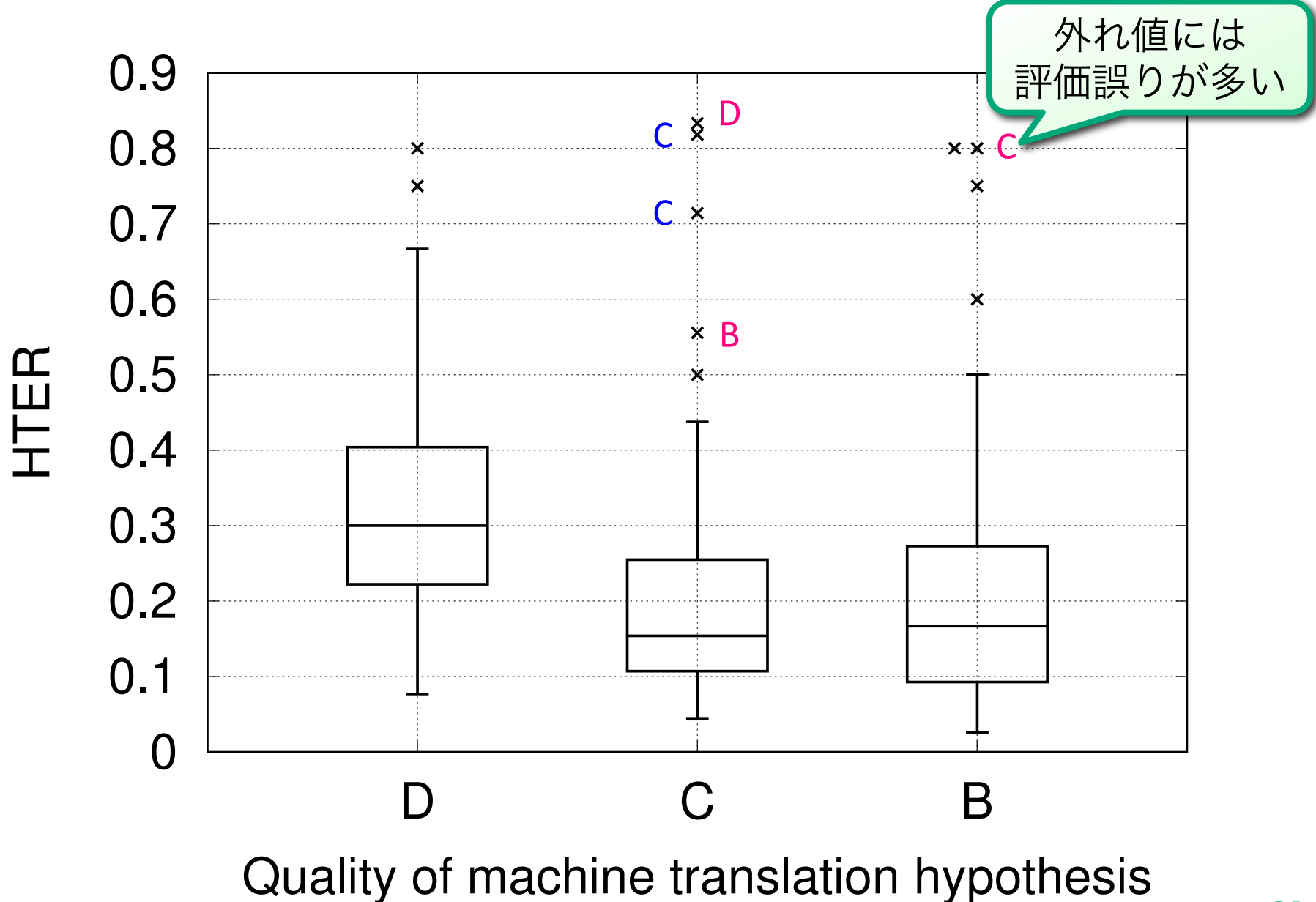
■ 否定辞の欠落

入力文	多額の現金は持ってこないでください。	
MT訳	Please bring a lot of cash.	評価D
修正訳	Please don't bring a lot of cash.	HTER=0.22

■ 指示対象の誤り

入力文	首が痛くありませんか。	
MT訳	Do you have pain in my neck?	評価D
修正訳	Do you have pain in your neck?	HTER=0.13

評価値とHTERの関係 (日韓, 病院内会話)



評価値がB/CだがHTERが高い例

■ 過度な編集/評価値も誤り？

入力文 一億二千六百五十三万人です。

MT訳 One hundred twenty six million five hundred thirty thousand.

修正訳 125,530,000 people. 評価C → B
HTER=3.00

入力文 手が痛いです。

MT訳 I have pain in my hand. 評価B → A

修正訳 My hand hurts. HTER=1.25

まとめ

■ MT訳の評価と修正訳を含む対訳コーパスの構築

- ねらい: MT品質の推定(QE) & MTユーザの満足度向上
- エントリの例

日本語原文	片道だけで買えますか。
模範訳	May I get it for one way?
MT訳 (new)	Can I buy just one way?
評価値 (new)	B
修正訳 (new)	Can I just buy a one way ticket ?

- 日本語 → 英語, 中国語, 韓国語
 - 日本語原文: 旅行会話8789発話, 病院内会話1,676発話
 - 句に基づくSMT (PBSMT)によるMT訳
- ALAGINを通じての公開を検討中