

翻訳時に参照すべき情報が 欠けることで生じる問題:

ニュース記事の英日機械翻訳・ポストエディットを例題に

藤田 篤



情報通信研究機構

今日の話題: MTの誤り分析

■ 機械翻訳(MT)(の当面の目標) vs. 翻訳

- 例題: ニュース記事の英日機械翻訳
- 分析の方法論
 - MT訳の2段階ポストエディット
 - MT訳 → MTの限界点 [Snover+ 06]
 - MTの限界点 → 最低限の翻訳
 - 2段目のポストエディット内容の分析
 - 人間翻訳向け校閲体系 [豊島+ 16] を用いた分類
 - MTが参照していない情報の確認

■ 提言

- MTの研究者や開発者向け
- MTのユーザ向け

翻訳とは何か?

■ 起点言語文書が担う役割を目標言語で果たすこと

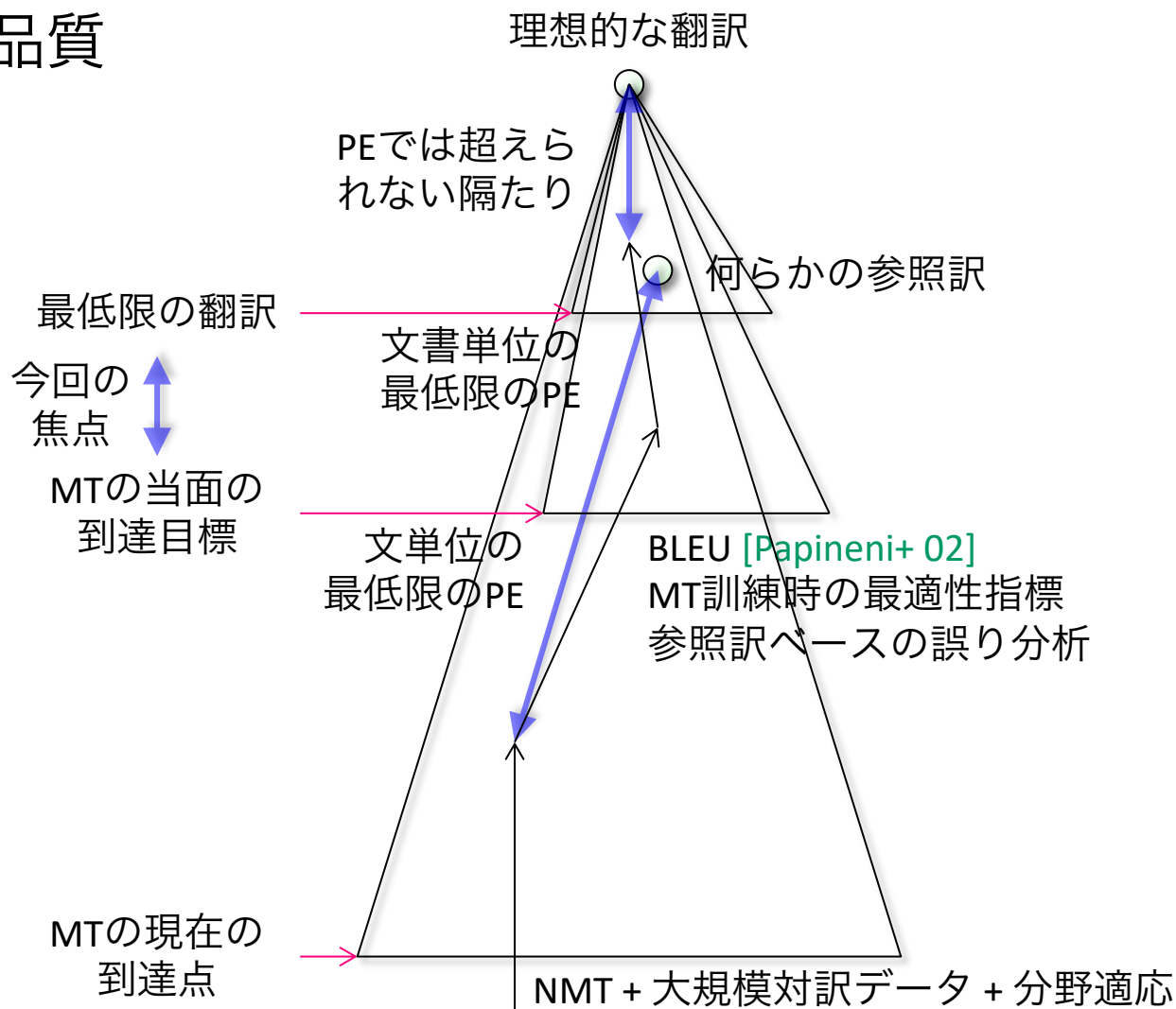
- 文書を扱う
 - 語や文の単なる sequence-to-sequence ではない
 - 文書の内部構造, 構成要素の役割, それらの相互作用
- 文書の外側にある仕様も要考慮
 - 文書の役割や翻訳の目的(スコプス) [Vermeer 89]
 - 目標言語における規範, 言語仕様域(レジスタ)

■ cf. MT

- 定式化のために単純化/無視
- 訓練データで間接的に表現

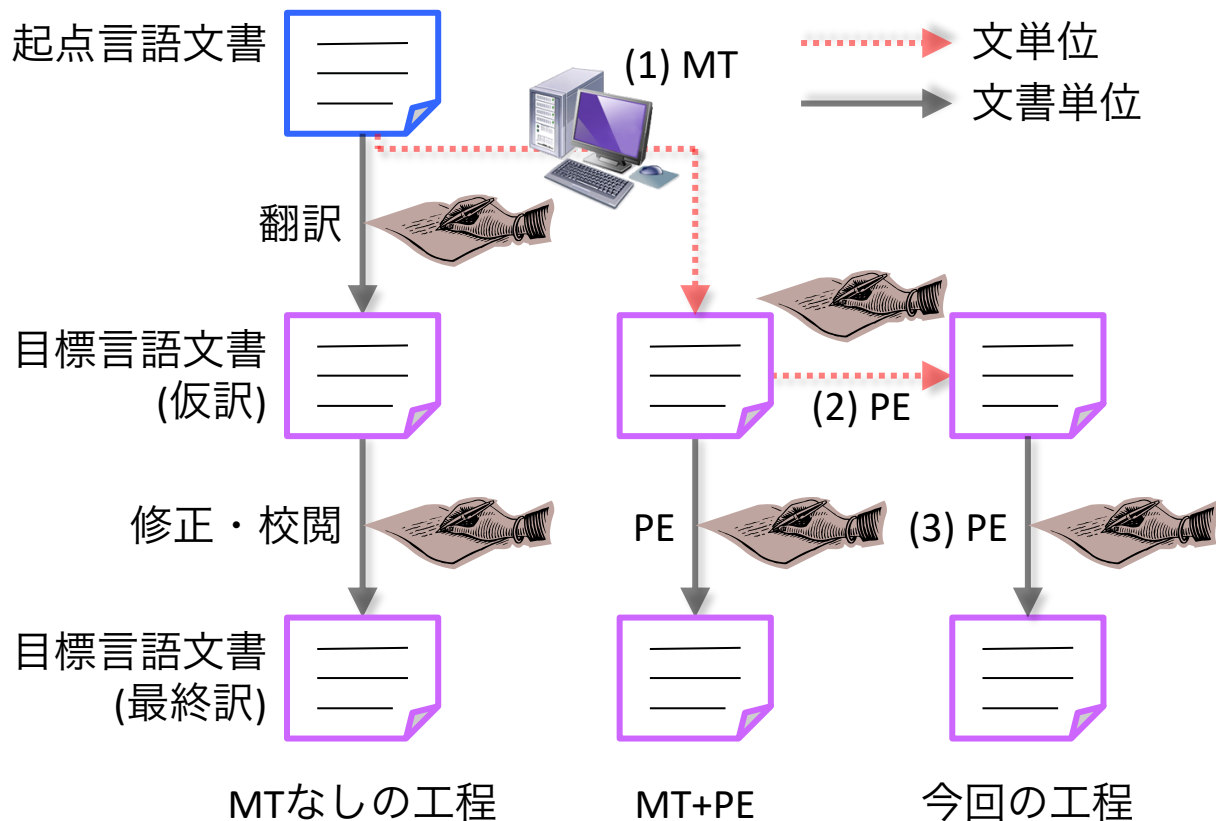
疑問とアプローチ

■ MTの到達目標の向こう側にどんな課題があるか？



分析対象データの作成

■ 翻訳工程の対比



分析対象文書

■ ニュース記事の英日翻訳

- 起点言語: それなりに分かる英語
- 目標言語: 母語=日本語
- その他の要求仕様
 - 原文が文書である
 - 訓練データがある: できるだけ高品質なMTから始めるため
 - 参照訳がある: PEの作業量の抑制のため

→ Asian Language Treebank [Riza+ 16]

用途	文書数	文数	トークン数	
			英語	日本語
訓練データ	1,698	18,088	472.1k	610.8k
開発データ	98	1,000	25.6k	33.0k
評価データ	97	1,018	26.3k	33.4k

工程(1) 文単位のMT訳の生成

■ 英日NMTの学習

- 詳細は予稿を参照されたい
- 3フェーズ [Chu+ 17]
 1. 事前学習 (NICT内製TexTraデータ)
 2. 混合パラメタ調整 (TexTraデータ + ALT訓練データ)
 3. 非混合パラメタ調整 (ALT訓練データ)

■ ALTの評価データを翻訳

- ビーム幅10のビームサーチ
- 長さに関する正規化重みは開発データを使って最適化

工程(2) 文単位のPE

- 全1,018文をシャッフル

- 文ごとに必要最低限の修正を施す [Snover+ 06]

- “最低限”: 模範訳よりはMT訳に似ている
 - 編集回数(MT訳 → 修正訳) $\leq$ 編集回数(MT訳 → 模範訳)
- 編集回数は tercom でカウント
 - 1語削除 = 1回
 - 1語挿入 = 1回
 - 1語置換 = 1回
 - ひとまとまりの語句の移動 = 1回

- 結果

- 95% (970/1,018) の文にPEが必要 → MTはまだまだ

工程(3) 文書単位のPE

■ シャッフルしてあった97文書の文を整列

■ 文書ごとに必要最低限の修正を施す

- 文ごとのときと同じ制約
- 結束性, 一貫性などに注意

■ 結果

- 31% (320/1,018) の文にPEが必要 → 文単位の翻訳の困難さ

訳出物	BLUE計算時の参照訳	
	ALTの参照訳	工程(3)の結果
ALTのみを用いた場合のMT訳	14.6	15.9
フェーズ1終了時のMT訳	29.0	40.3
フェーズ2終了時のMT訳	35.8	57.6
フェーズ3終了時のMT訳	36.0	58.6
工程(2): 文単位のPEの結果	36.8	95.0
工程(3): 文書単位のPEの結果	36.8	100.0

PE事例 (ID=139760)

原文

John Paul's six-day tour was hugely popular, attracting hundreds of thousands of people from the Catholic religion. It became the first ever UK visit from the Pope, which will make Pope Benedict's visit the second papal one.

MT訳

工程(1) 文単位のMT訳の生成

ジョン・ポールの6日間のツアーは非常に人気があって、カトリックの宗教から何十万もの人々を引きつけた。それはローマ教皇からの初めての英国訪問となり、ローマ教皇ベネディクトの第二の教皇訪問となる。

文単位のPEの結果

工程(2) 文単位のPE

ジョン・ポールの6日間のツアーは非常に人気があり、カトリックを信仰する何十万もの人々を引きつけた。それはローマ教皇の初めての英国訪問となり、ローマ教皇ベネディクトの訪問は2度目の教皇訪問となる。

文書単位のPEの結果

工程(3) 文書単位のPE

法王ヨハネ・パウロの6日間の訪問は非常に人気があり、カトリックを信仰する何十万もの人々を引きつけた。この時がローマ法王の初めての英国訪問となり、法王ベネディクトの訪問は2度目の法王訪問となる。

誤り分析

■ MTの限界点と最低限の翻訳の差を知りたい

- 工程(3) 文書単位のPEの内容を分析する

■ 手順

1. 事例の列挙: 修正前後の表現対の抽出・分解 [宮田+ 17]
 - 原文の対応箇所も見ながら
2. 事例の分類: MNH-TTの校閲カテゴリ体系 [豊島+ 16]
 - 人間の英日翻訳向けに設計されたもの, 16カテゴリ
3. 文単位での修正可否の判定
 - = 文書中の他の文を参照する必要の有無
 - 工程(2) 文単位PEの目標 vs. 文単位の最適訳

修正事例の列挙・分解・分類の例

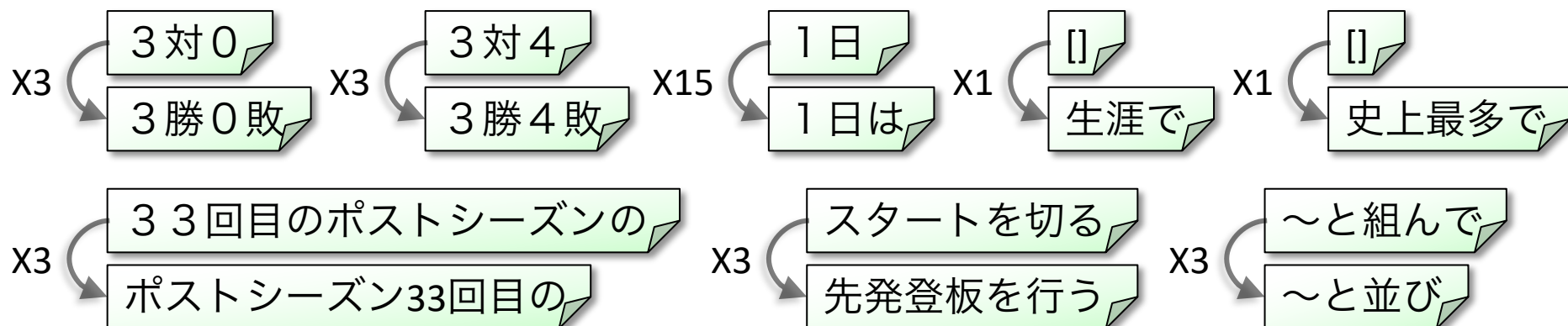
文単位のPEの結果

クレメンス（ワールドシリーズ7回出場で3対0、防御率1.90）は、少なくとも1日ペティット（ワールドシリーズ10回出場で3対4、3.90）と組んで、土曜日に33回目のポストシーズンのスタートを切る。

文書単位のPEの結果

工程(3): 文書単位のPE

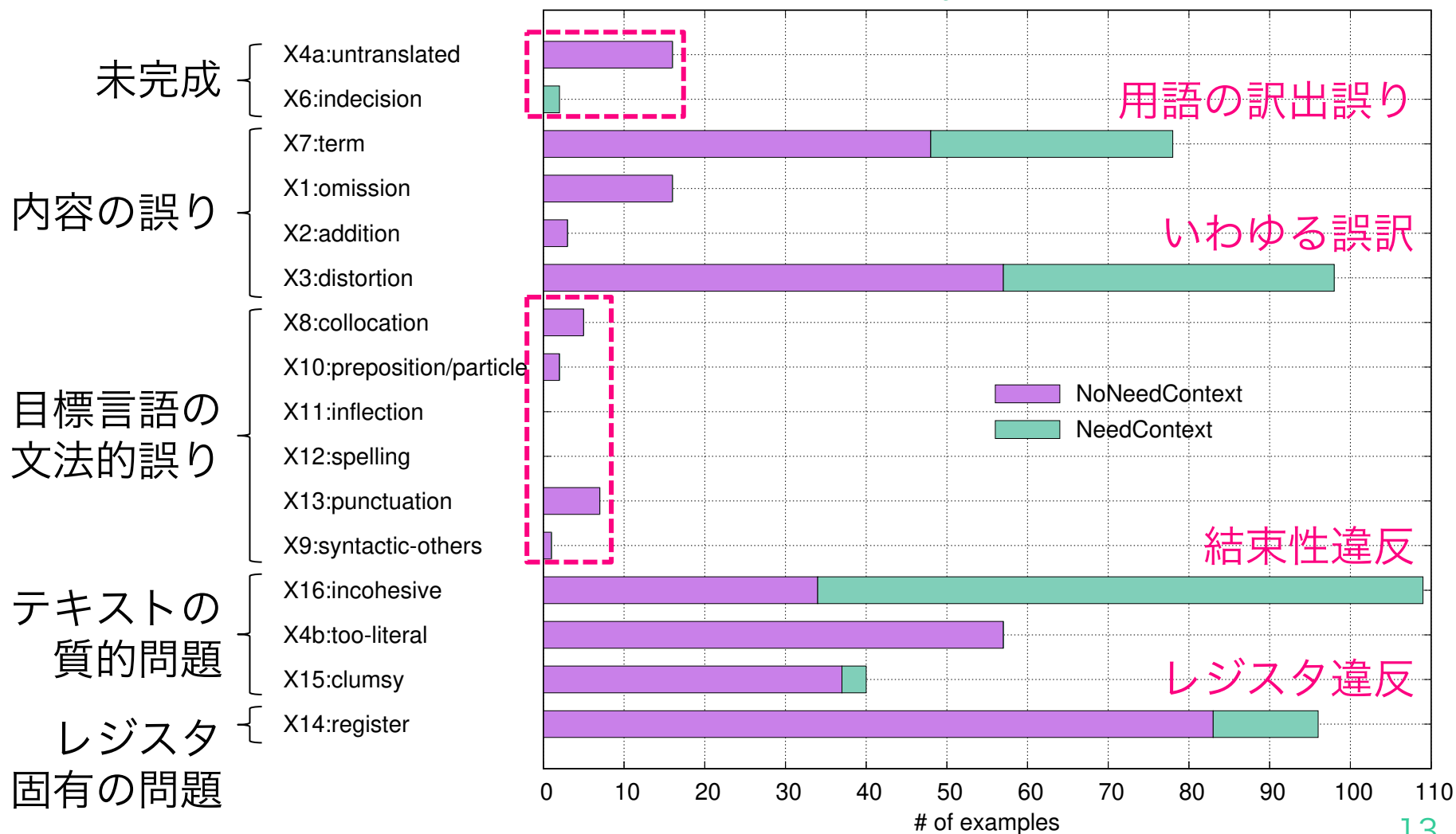
クレメンス（ワールドシリーズ7回出場で3勝0敗、防御率1.90）は、少なくとも1日はペティット（ワールドシリーズ10回出場で3勝4敗、3.90）と史上最多で並び、土曜日に生涯でポストシーズン33回目の先発登板を行う。



各カテゴリの事例数 (合計530件)

■ 文書中の他の文を参照する必要の有無で細分類

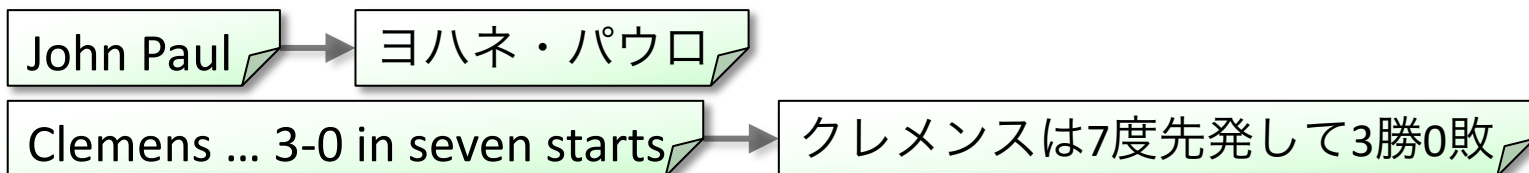
- cf. 人間の翻訳 [豊島+ 16][山本+ 16][Fujita+ 17][田中+ 18]



文単位でもできることはまだ多い

■ ただし、単なる sequence-to-sequence ではない

- **知識**が不可欠: 一般常識, 対象分野の専門的知識や表現



- **翻訳の仕様**を要理解
 - スコパス: 文書の役割や翻訳の目的
 - e.g., 想定読者, 文書の用途, メディアの空間的制約
 - レジスタの規範
 - e.g., 固有名の翻字の要不要, 使用可能な漢字, 括弧の使い方
- **用語集**を要参照

文書内の他の文を参照する必要性

■ 修正事例中の31% (164/530)

- 【X16 結束性違反】 が最多の75件
 - 明示化: 指示詞や人称代名詞等の参照先の名詞句の明示
 - 省略: 自明な格要素の削除
 - 主題化/非主題化: 「は」と「が」の使い分け
 - 訳出の統一: 固有名は【X7 用語の訳出誤り】に分類
- 文脈依存の曖昧性の適切な解消

The United Nations (UN) on Friday called on the government of Pakistan to investigate numerous killings and abductions, particularly of journalists. ... Amnesty International has **also urged** the government of Pakistan to ensure the protection of journalists and media workers in the country.

Since completion of the surgery, the patient has been on immuno-suppressant drugs to prevent rejection of the new **limbs**.

提言: MTの研究者や開発者向け

■ 何を目指すか?

- 全自動? 人間との共存?
- 文単位でも問題は山積
 - 工程(2) 文単位のPEで95%が修正された
 - 工程(3) 文書単位のPEにおける各文で可能な修正
- 陽に参照・遵守していない情報
 - 知識, 翻訳の仕様, 用語集
 - 一部は制約付きデコーディング [Post+ 18] に期待
 - 文脈情報
 - e.g., ゼロ照応, 指示詞, 上位-下位, オブジェクト-属性

■ 文書レベルの自動評価技術を要確立

- 訳出の一貫性, 結束性など

提言: MTのユーザ向け

■ 現在のMTは必ずしも翻訳をしない

- 翻訳時に参照すべき情報の多くを見ていない
- 単なる sequence-to-sequence

■ そのようなMTを利用する上での注意

- 意中のMTが何を参照しているかを知る
- 使い方をよく考える
 - 翻訳の目的や仕様などに照らして
 - 誰が消費する? 誤りの責任は誰が負う?
- さもなくば, 深刻な翻訳過誤を生じうる

堺筋線

Sakai Muscle line

高塚川周辺に避難勧告が出ました

Foi emitido **aviso para se refugiar** proximidades do rio Takatsuka.

まとめ

■ 機械翻訳(MT)(の当面の目標) vs. 翻訳

- 例題: ニュース記事の英日機械翻訳
- 分析の方法論
 - MT訳の2段階ポストエディット
 - MT訳 → MTの限界点 [Snover+ 06]
 - MTの限界点 → 最低限の翻訳
 - 2段目のポストエディット内容の分析
 - 人間翻訳向け校閲体系 [豊島+ 16] を用いた分類
 - MTが参照していない情報の確認

■ 提言

- MTの研究者や開発者向け
- MTのユーザ向け