Attainable Text-to-text MT vs. Translation: Issues Beyond Linguistic Processing

Atsushi Fujita

NICT, Japan

Given a document, how would you translate it?

…

So the following statement should have been taken into account, among others:

> *States Parties particularly condemn racial segregation and apartheid and undertake to prevent, prohibit and eradicate all practices of this nature in territories under their jurisdiction.*

Justice Baltasar Garzón decided to indict Augusto Pinochet in 1998, he was

…

Given a document, how would you translate it?

Register? Text type? Readers?

Purpose? Usages?

…

So the following statement should have been taken into account, among others:

States Parties particularly condemn racial segregation and apartheid and undertake to prevent, prohibit and eradicate all practices of this nature in territories under their jurisdiction.

Justice Baltasar Garzón decided to indict Augusto Pinochet in 1998, he was

…

Quote from an international treaty

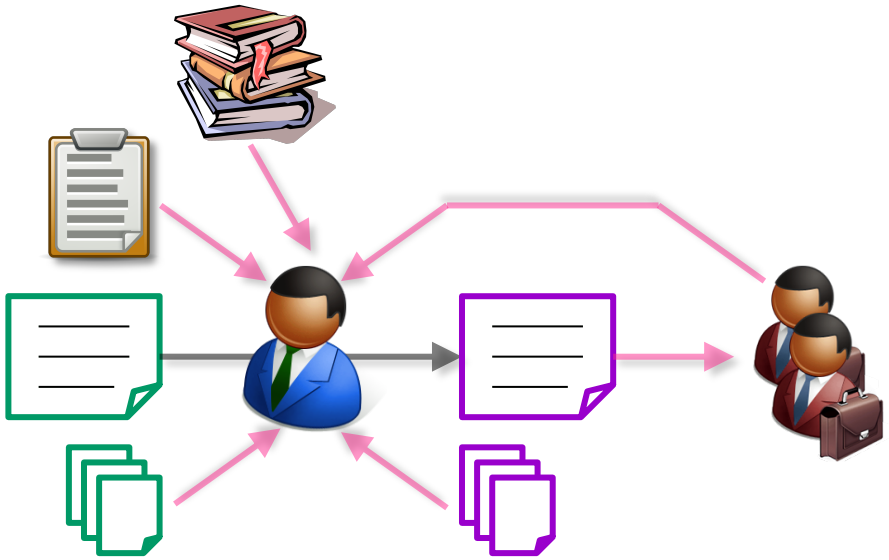
Does official translation exist?

Person names

Existing transliteration?

Translation

- Producing a document in the target language that plays the same role as the given document
- The production workflow refers to
  - Not only documents
    - Text segments (as in MT)
    - Other modals within the given document
  - Extra-document and non-linguistic information
    - Designated specifications [ISO/TC37 15]
    - Norm [Toury 78]
      - Conventions and criteria in the target language and register
      - Register: determined by the subject field, expected readers, mode, etc.
    - Skopos [Vermeer 89/04]: objective and intended usages of final product



Our aim

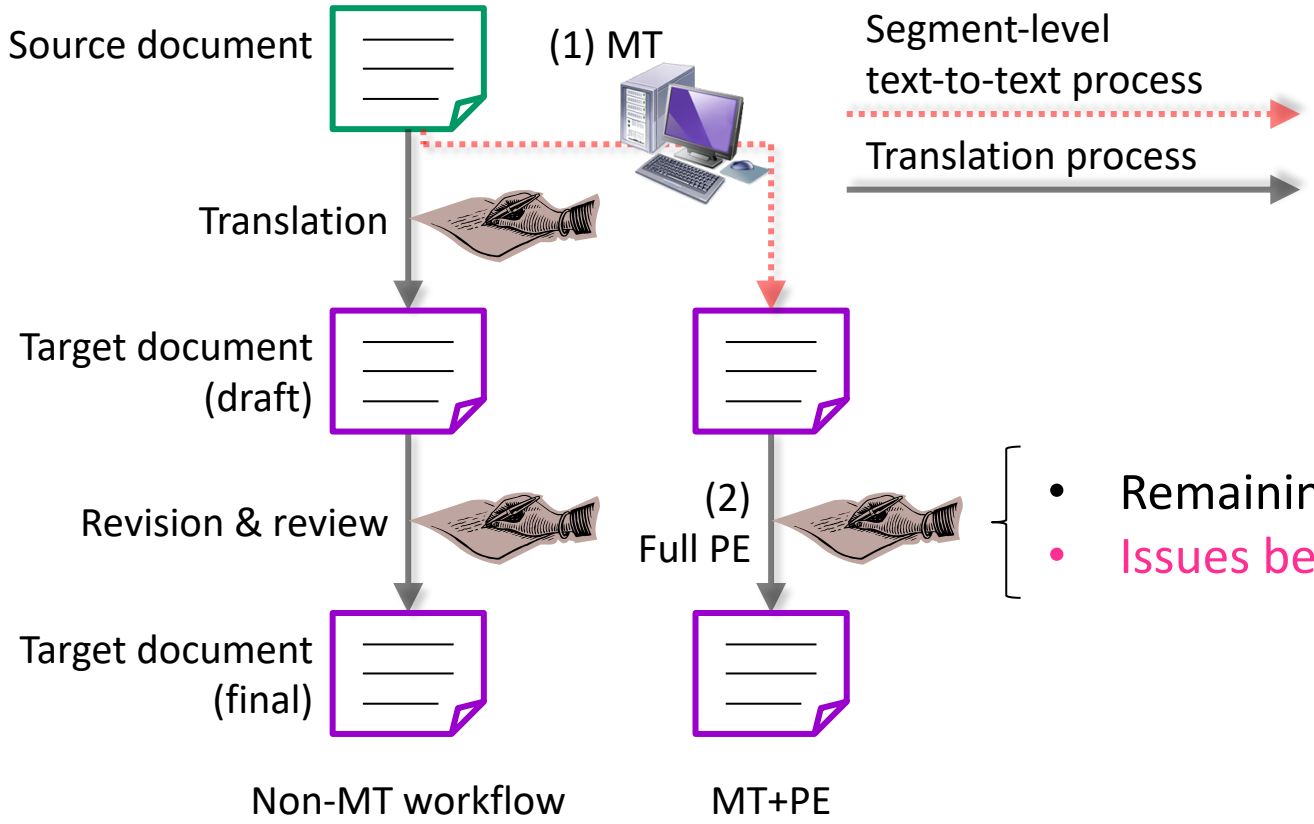
- Clarify the gap between MT and translation
  - through analyzing translation issues that MT systems cause
  - (1) Collecting issues and their revisions through PE
  - (2) Clarifying the types of necessary information

- Discuss
  - Future research directions
  - How to properly reflect necessary information in MT

Collecting and analyzing revision examples of translation issues

A strategy to fill the gap between MT and translation

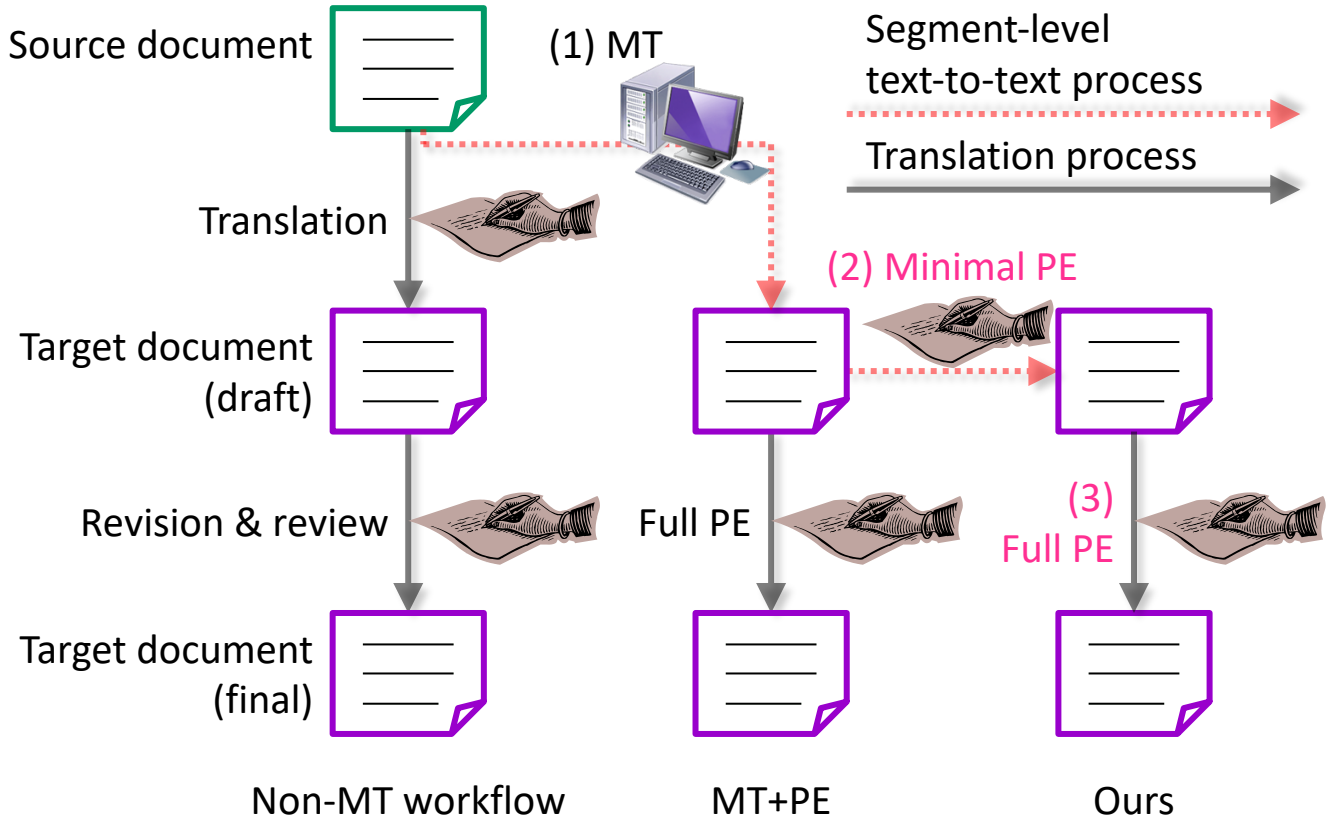
- MT+PE workflow
  - ISO 18587 [ISO/TC37 17]



- Remaining issues of text-to-text MT
- Issues beyond text-to-text MT

Our 2-stage PE workflow

- Given MT outputs (1), we perform
  - (2) Segment-level minimal PE: referring only to each segment, only necessary editing
  - (3) Document-level full PE: referring to any information, perform all necessary editing



(1) Segment-level text-to-text MT

- A case study: English-to-Japanese news translation task
  - Asian Language Treebank [Riza+ 16]
  - English-to-Japanese NMT system
    - Transformer Base [Vaswani+ 17]
    - Trained via a 3-phase domain adaptation [Chu+ 17][Dabre+ 19]
      1. Pre-training (gigantic in-house data)
      2. Mixed fine-tuning (gigantic in-house data + 18k ALT training data)
      3. Pure fine-tuning (18k ALT training data)
  - Starting with translations with reasonably high BLEU score
  - See the paper for details

ALT-Standard-Split	#Doc.	#Seg.	#Tok.	
			En	Ja
Training	1,698	18,088	472.1k	610.8k
Development	98	1,000	25.6k	33.0k
Test	97	1,018	26.3k	33.4k

Translation	BLEU	HTER
Output of NMT trained only on ALT	14.6	73.9
Output of NMT in phase 1	29.0	55.5
Output of NMT in phase 2	35.8	47.6
Output of NMT in phase 3	36.0	47.6

(2) Segment-level minimal PE

- Shuffle all the 1,018 segments in the ALT test data
- Perform minimal PE for each segment
  - at an experienced, ISO-certified TSP [ISO/TC37, 15;17]
  - referring only to the segment itself
  - "minimal PE" (cf. [Scarton+ 15])
    - TER(***mt*** → ***pe***) ≦ TER(***mt*** → ***ref***)
      - ***mt***: MT output, ***pe***: PE result, ***ref***: reference translation
      - Restart from ***mt*** if ***pe*** does not satisfy the constraint
    - Prohibit subjective stylistic changes
- Results
  - 95% (970/1,018) of the segments were revised
  - The closest goal of segment-level text-to-text MT

(3) Document-level full PE

- Reorder the text segments into the original 97 documents
- Perform full PE for each document
  - at an experienced, ISO-certified TSP [ISO/TC37, 15;17]
  - as usual MT+PE workflow
    - regarding the segment-level PE results as the raw MT output
- Results
  - 31% (320/1,018) of the segments were revised

An example (DocID=139760, SntID=6023)

Source

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.

(1) Segment-level text-to-text MT

MT output

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。

(2) Segment-level minimal PE

Result of segment-level minimal 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。

(3) Document-level full PE

Result of document-level full 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。

Error analysis

- Objective: clarify the types of necessary information
- Procedure
  - Extract revision examples produced during stage (3)
    - Identify the corresponding text spans, also referring to the source text
    - 529 examples
  - Annotate each example (an annotator ≠ the PE worker)
    - Need for document-level textual information
    - Need for extra information
      - and the type(s) of such information
    - Issue type among 16 types [Fujita+ 17]

A simplified example (Figure 2 in the paper)

Source

Clemens (3-0, 1.90 ERA in seven World Series starts) will make his 33rd career postseason start Saturday, at least for a day matching Pettitte for the most ever.

(1) Segment-level MT & (2) Segment-level PE

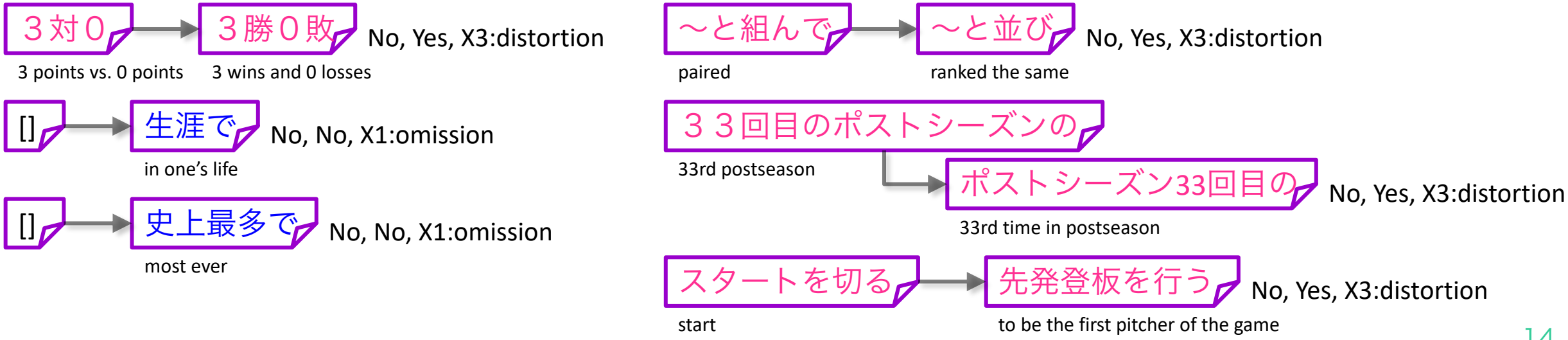
Result of segment-level minimal PE

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日はペティットと組んで、土曜日に３３回目のポストシーズンのスタートを切る。

(3) Document-level full PE

Result of document-level full PE

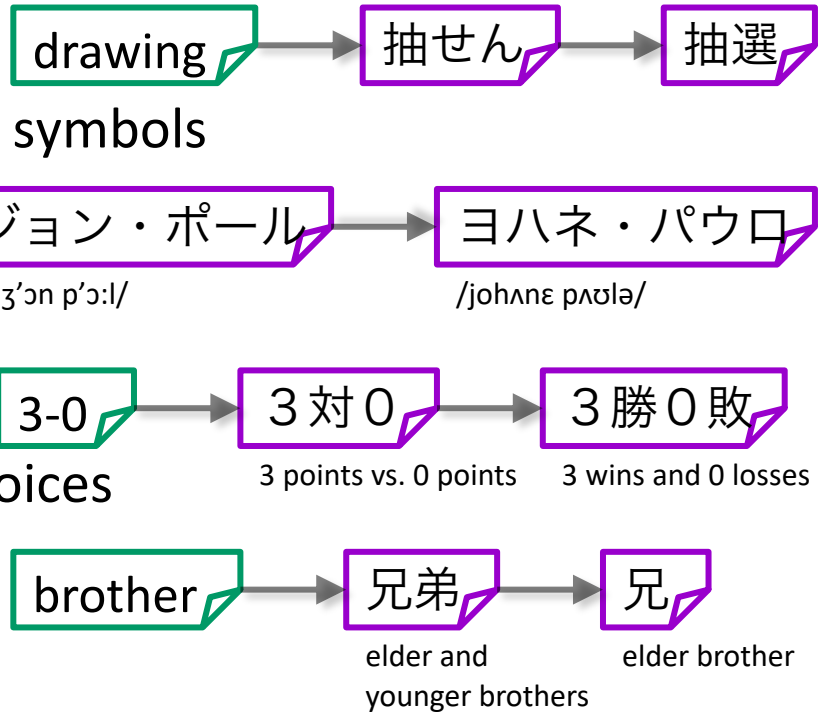
クレメンス（ワールドシリーズ７回出場で３勝０敗、防御率１．９０）は、少なくとも１日はペティットと史上最多で並び、土曜日に生涯でポストシーズン33回目の先発登板を行う。



Necessity of extra information

Necessary info.		Extra info.	
		No need	Necessary
Document-level textual info.	No need	196	168
	Necessary	116	49

- 40% (217/529) of revisions relied on extra information
  - (A) Fine-grained style specifications (121/217)
    - Violation of circumstances: vocabulary, character, usages of symbols
  - (B) Terminology (80/217)
    - Incorrect term translations, untranslated Latin names
  - (C) Domain-specific knowledge (31/217)
    - Misunderstandings of the contents, inappropriate lexical choices
  - (D) Reference documents (8/217)
    - Wrong disambiguations, remaining ambiguities



Discussion on future research directions

What is the appropriate goal of MT (and researchers)

- Full automatic MT is infeasible [Hutchins+ 92]
- Human-Machine cooperation is vital (e.g., MT+PE)
  - Humans (users of MT)
    - Determine the best practices in using MT
    - Acquire translation literacy [Klitgård 18] and MT literacy [Bowker+ 19]
      - Our data could be a useful learning material
  - Machine (to be used by humans in production workflows)
    - Resolve the remaining issues of text-to-text MT
      - 60% of document-level PE rely only on text
    - Enable MT conform to the given criteria
    - Enable MT confirm the entities and factuality

Toward properly reflecting extra information

- How to realize perfect adherence to the given criteria?
  - A variety of rules and constraints
    - Constrained decoding
      - Vocabulary filtering [Ha+ 17]
      - Terminology guided decoding [Hasler+ 18, Post+ 18, Zhang+ 18]
    - Post-processing
      - e.g., rules/pattern in CAT tools
  - cf. Indirect supervision through corpus-level adaptation
    - Domain adaptation [Chu+ 17]
    - Style transfer [Niu+ 17, Michel+ 18]

大項目	中項目	小項目	
基本文型 (phrase style)	文体	本文	敬体
		見出し	体言止め （動作を説明する見出し:「～の～」ではなく「～を～する」で統一）
		操作手順	原則的に「ですます調」を使用。ユーザーに特別な操作を指示する場合を除き、「～してください。」は使用しない。
		箇条書き	原則として体言止め 箇条書きの項目が文の場合は、原則的に「ですます調」
		図表内テキスト	表中の見出しは体言止めを使用し、項目が文の場合は「ですます調」を使用
		図表のキャプション	原則的に英文が名詞句の場合は名詞句、文の場合は「ですます調」を使用
文字の表記 (ideographics)	用字、用語	漢字	原則として常用漢字を使用。
		漢字の送りがな・複合語の送りがな	昭和48年6月18日内閣告示第2号「送り仮名の付け方」に準じる。
		カタカナの長音	原則として、末尾の長音も含めて文字数が4文字以下のときは、そのまま長音符号は残す。末尾の長音も含めて文字数が5文字以上の場合は、末尾の長音符号を削除する。 カタカナ複合語（原文の英単語が2語以上のもの）は、それぞれの単語に区切ってカウントして判断する。
		算用数字と漢数字の使い分け	原則的に、数えられる語句に算用数字（半角）を使用し、漢数字が一般的である場合は漢数字を使用
		一部の助数詞の表記	「～か月」、「～か所」
文字間のスペース (spacing)	単一文字間のスペースの有無	全角と半角の間	スペースを入れない
		全角どうし	スペースを入れない
		半角どうし	和文中に英文を引用するなど、和文に英文が含まれる場合は英文中の半角スペースを維持する
	カタカナ語間のスペースの有無	カタカナ複合語	半角スペースおよび中黒を入れない
記号の表記と用途	記号 (symbols)	句読点	全角の「。」および「、」を使用する
		感嘆符(!)	半角
		疑問符(?)	半角
		スラッシュ(/)	半角
		波線(～)	全角
		コロン(:)	全角
		引用符('または")	日本語訳文では基本的に使用しない。原文にある場合は「」で置き換える
		数式記号 (+-*/)	半角
		パーセント (%)	半角
	かっこ (parentheses)	丸かっこ ()	全角
		大かっこ []	半角
		かぎかっこ「」	全角
		二重かぎかっこ『』	全角
単位の表記 (quantity units)	単位記号の表記	単位	数字と単位の間に半角スペースを入れない
		通貨	翻字

Conclusion and future work

- Error analysis of NMT
  - Manual PE for filling the gap between MT and translation
  - Types of necessary information

- Discussion
  - Future research directions
  - How to properly reflect necessary information?

- Future & ongoing work
  - Evaluate the impact of enforcing decoding
  - Compile educational materials