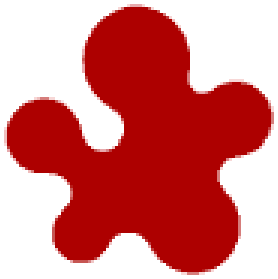
NLP若手の会第4回シンポジウム

# プレゼンテーションの構成要素



公立はこだて未来大学

2009年9月30日

# スライド翻訳における分類と前後処理

○野口 耕自朗，　藤田 篤，　松原 仁

# 概要

- 既存の翻訳器を用いた発表用スライド翻訳
- スライド翻訳における構成要素分類と処理
  - スライド中の表現を構成要素に分類
  - 構成要素の種類に応じて翻訳すべきか否か判定
  - 構成要素の種類ごとに判定方法および翻訳の前後処理を検討

# 背景と目的

## 背景

- 異文化コラボレーションの機会増加
  - グローバル化，ネットワークの発達
- 言語の壁

## 目的

- 発表用スライド翻訳における構成要素の分類
- 要素の種類に応じた処理手法による精度

良い翻訳の実現

# スライド翻訳例

- 引用表現の翻訳 [inyou,2009]
  - 複合名詞誤訳多発
- 体言止めの使用が頻繁
  - *(強調表現の消失)*

例)「サーバへアクセス」

- 今回の調査目的
  - 大規模データを使用⇒訳質向上
  - 10億件のラベル付きデータ<x,y>を使用

現在の翻訳器の翻訳

# A slide translation example

- Translation [ i n y o u of quote representation.2 0 0 9 ]
- Compound noun mistranslation frequent occurrence
- Use of substantive end-form is frequent.
- (Disappearance of an emphasis expression)

An example) "I access a server."

- This investigation purpose
- Large-scale data is used,⇒ translations matter improvement
- 1 Data < x with 0 labels.y > is used.

引用の[]が誤訳

翻訳したくない例文が翻訳されてしまう

「⇒」の前後が分割されて翻訳されない

数値，数式交じりの文が誤訳

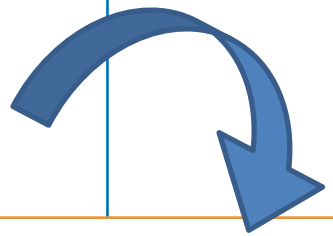
# スライド翻訳例

- 引用表現の翻訳 [inyou,2009]
  - 複合名詞誤訳多発
- 体言止めの使用が頻繁
  - *(強調表現の消失)*

例)「サーバへアクセス」

- 今回の調査目的
  - 大規模データを使用⇒訳質向上
  - 10億件のラベル付きデータ$<x,y>$を使用

# 理想的な翻訳



# An example of slide translation

- Translation of quote expression [inyou,2009]
  - Compound nouns are frequently mistranslated.
- Substantive end-forms are frequently used.
  - *(Disappearance of emphasis expressions)*

ex)「サーバへアクセス」

- Aim of this research
  - Use large-scale data ⇒ Translation quality improves
  - 1 billion labeled data $<x,y>$ are used

# アプローチ

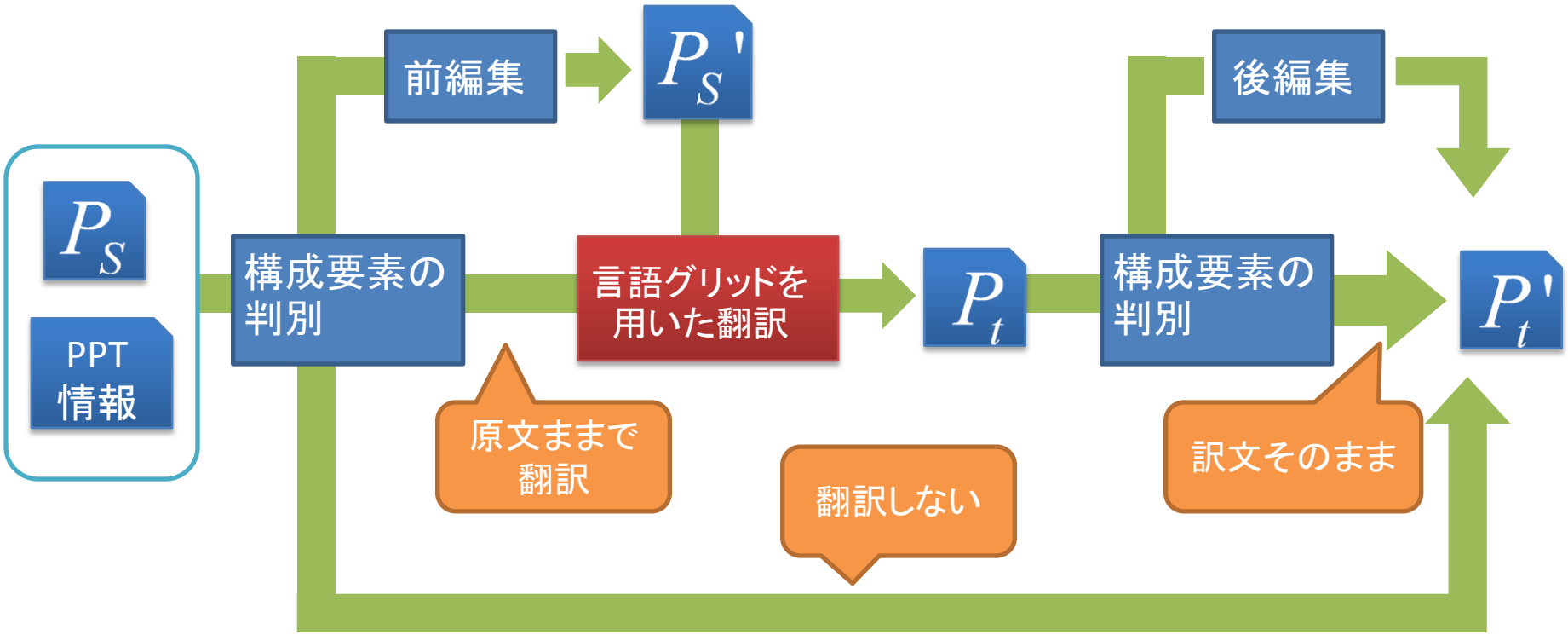
- スライドの構成要素に着目した翻訳
  - スライド構成要素に応じて翻訳方法を選択
    - スライド構成要素の判別方法
    - スライド構成要素に対する前後処理

- スライド翻訳
  - 対象：Microsoft PowerPointファイル
  - 対象言語：「日→＊」
  - 翻訳には既存の翻訳器を使用
    - 言語グリッド [Ishida,2006]

[Ishida 06] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration,
in Proc. of IEEE/IPSJ Symposium on Application and the Internet (SAINT–06), pp.96–100 (2006).

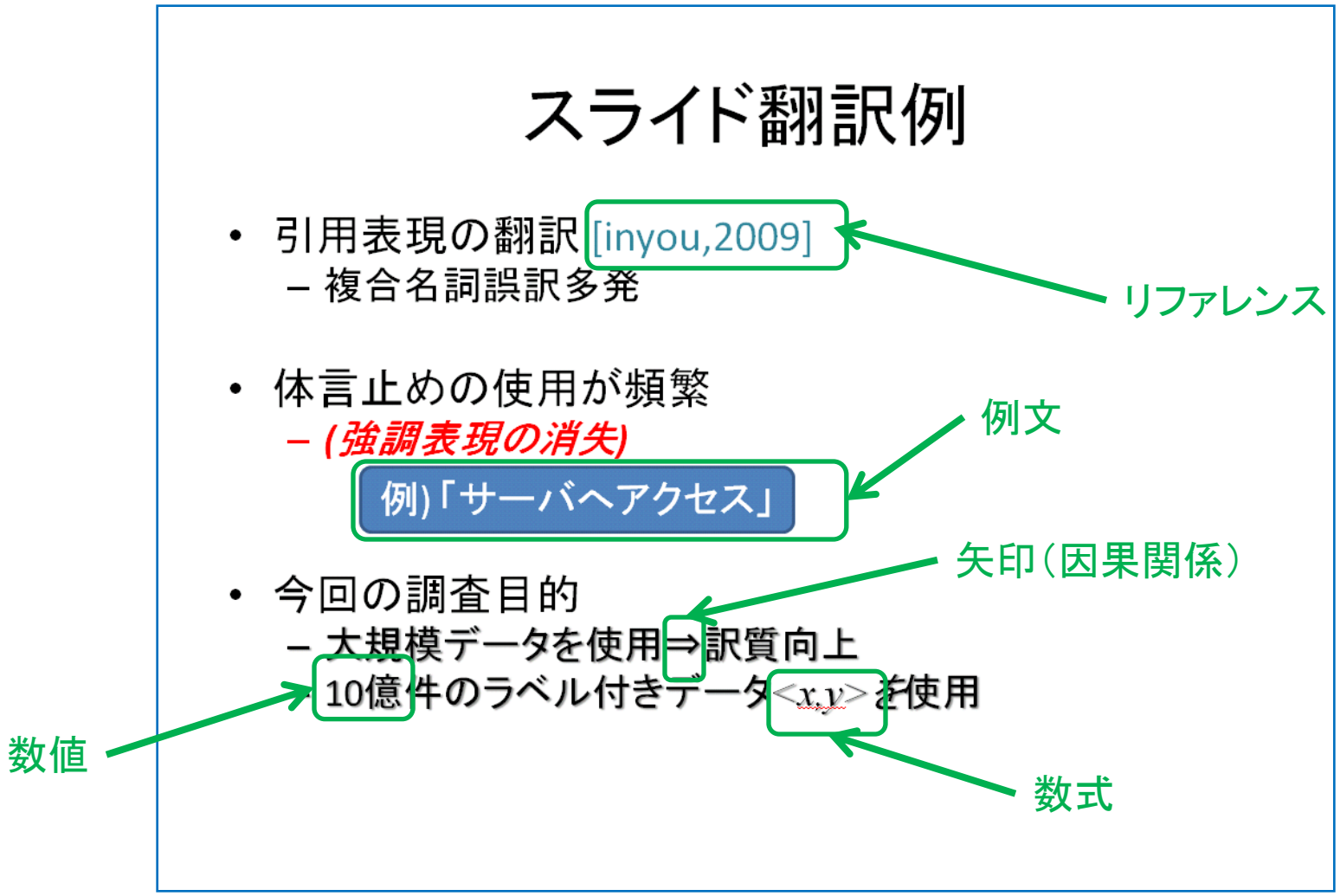
# 前後処理のフロー

- 構成要素に応じて前後処理を切り替え




Ps = 翻訳元言語の表現
Pt = 翻訳先言語の表現

# スライド構成要素の例


スライド翻訳例
• 引用表現の翻訳 [inyou,2009]
– 複合名詞誤訳多発
• 体言止めの使用が頻繁
– (強調表現の消失)
例)「サーバへアクセス」
• 今回の調査目的
– 大規模データを使用⇒訳質向上
10億件のラベル付きデータ<x,y>を使用
リファレンス
例文
矢印(因果関係)
数値
数式

# スライド構成要素の分類と対処(1/2)

| スライド構成要素 | 翻訳処理 |
|---|---|
| ・筆者情報 | カンマやスペースで区切って要素ごとに翻訳. |
| ・文章 | そのまま翻訳. |
| ・専門用語 | 専門用語辞書の作成が必要. |
| ・複合名詞 | CaboChaで構文解析後，接続詞を添付．接続詞の添付候補を単語ごとに変化. |
| ・体言止め表現 | 「それは～である」の形に変換して翻訳．その後，追加した部分に当たる「It's」や「That」を消去. |
| ・単体の例文 | そのまま．「例」「ex」などをキーワードとして直後の文を例文と認識．また，意味内容の明らかに違う文章を例文として認識 |
| ・陳述内の例文 | そのまま．意味内容の明らかに違う文章を例文として認識．「例」「例えば「-例文-」～」などの形があれば利用する. |
| ・引用文 | そのまま翻訳. |
| ・リファレンス | [ ]を手掛かりに切り出し，翻訳後に埋め込みなおす．カッコの内容が「年数＋人名（らしき文字列）」ならば未翻訳．それ以外ならカッコ内も翻訳する. |
| ・数式，変数入り括弧書き(),<> | ()＜＞を手掛かりに切り出し，翻訳後に埋め込みなおす．カッコの内容が「記号，ローマ字」の構成ならば未翻訳. |

# スライド構成要素の分類と対処(2/2)

| スライド構成要素 | 翻訳処理 |
|---|---|
| ・文章，単語入り括弧書き(),<> | ()<>を規準に切り出す．カッコ内が例文や引用，数式の条件に当てはまらない場合，カッコ内も翻訳する． |
| ・数式 | 記号，数字，演算子で構成されたものを数式と認識 |
| ・模式 | 演算子の前後で分割してそれぞれ翻訳 |
| ・数値 | 数値の羅列を判定 |
| ・記号 | 記号の自体はそのまま．<br>記号の前後の表現に対して記号に応じた処理を行う． |
| ・音訳 | ローマ字の羅列に対して，「ローマ字⇒日本語変換」をした後，日本語と英語を判別． |
| ・矢印（処理） | ひとまずそのまま翻訳． |
| ・矢印（前後の文）・結論 | 矢印前後の2文を「～は～」の形に結合して翻訳する．翻訳後に2文に分割． |
| ・矢印（前後の文）・因果関係 | 矢印前後の2文を「～のため～となる」の形に結合して翻訳する．翻訳後に2文に分割． |
| ・矢印（前後の文）・順番 | そのまま翻訳． |
| ・図形内文章（説明） | そのまま翻訳 |
| ・図形内文章（セリフ） | そのまま翻訳． |

# スライド翻訳例

- 引用表現の翻訳 [inyou,2009]
  - 複合名詞誤訳多発
- 体言止めの使用が頻繁
  - *(強調表現の消失)*

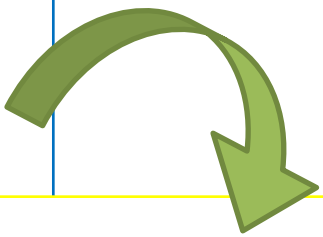
例)「サーバへアクセス」

- 今回の調査目的
  - 大規模データを使用⇒訳質向上
  - 10億件のラベル付きデータ<x,y>を使用

分類・処理の評価

レイアウトの評価

# 現在のシステムによる翻訳



## A slide translation example

- Translation of quote representation [inyou,2009]
- Compound noun mistranslation frequent occurrence
- Use of substantive end-form is frequent.
- (Disappearance of an emphasis expression)

An example) "I access a server."

- This investigation purpose
- Large-scale data is used.⇒.Translations matter improvement
- Data <x,y> use with 1000,000,000 sets of label.

# 表示情報の分類と対処

- レイアウト情報の分類

| 分類 | 処理 |
|---|---|
| インデント | 上位階層の情報を保持して翻訳 |
| 色つき（強調） | 翻訳後，原文との単語の対応により該当箇所を判定し，色づけ |
| 変形（ボールド等） | 翻訳後，原文との単語の対応により該当箇所を判定し，変形 |
| フォント | そのまま |

- 言語的表現の分類：
  - 各表現の持つ特性を翻訳後へ反映
    - 名詞句の命令文への誤訳を防ぐ
    - 受動形と能動形

# 関連研究

- 電子会議システムにおけるスライド翻訳
[宗森ら，2005]
  - スライド翻訳に関しては宗森らの研究を参考

- スライド頻出表現の翻訳前処理 [石黒ら，2009]
  - 連体節を含む文・・・「『名詞』の『名詞』」の形に換言
  - 主語のない文・・・仮主語の補完と削除

[宗森ら,2005] 宗森純，重信智宏，丸野普治，尾崎裕史，大野純佳，吉野孝.
“異文化コラボレーションへのマルチメディア電子会議システムの適用とその効果”. 情報処理学会論文誌，Vol. 46，No. 1，pp. 26-37，2005.

[石黒ら,2009] 石黒 雄佑,Villavicencio Paul,花植 康一,渡邉 豊英. ″言語グリッドを用いたプレゼンテーションスライドの日英翻訳の試み“.
IEICE technical report,**人工知能と知識処理研究会** 108(441) pp.73－78, 2009.

# 今後の計画

- 構成要素判別方法の検討
  - 構成要素の自動判別
    - PPT操作用APIによる判定
    - 正規表現による判定
  - 人手での構成要素の指定

- 評価
  - 評価基準となるスライド集合を事前に人手で翻訳
    - 「日→英」翻訳に関して評価
    - 英語ネイティブに翻訳を依頼
  - 2種類の方法で評価
    - 自動評価法「BLEU」
    - 人手で翻訳結果を評価