

# 産業文書に関する 言語処理技術の開発と期待

豊橋技術科学大学

井佐原均

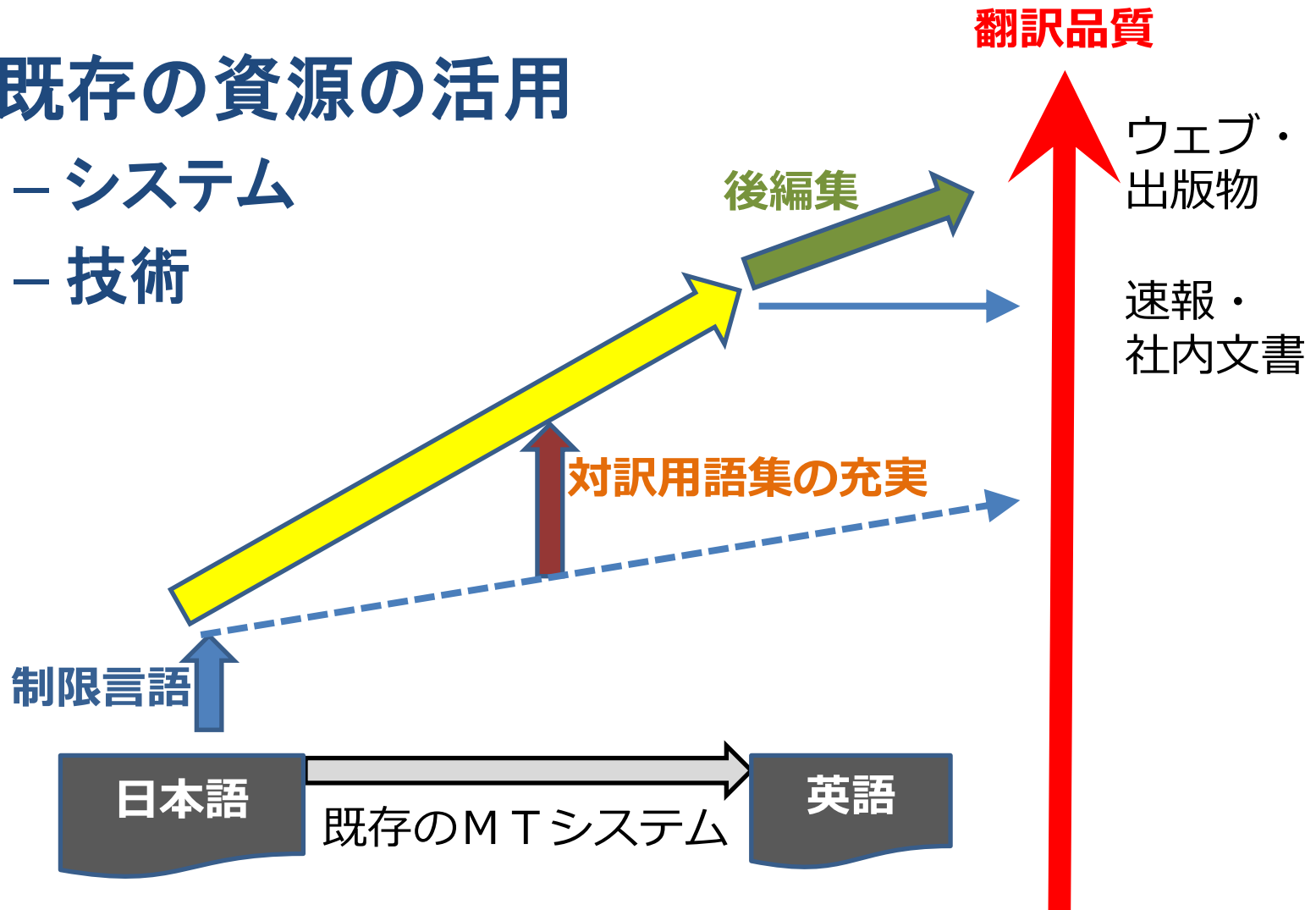
# より広い連携を目指して

- 産業日本語研究会（世話人会）
- テクニカルコミュニケーター協会
  - 高橋尚子専務理事
  - テクニカルコミュニケーションシンポジウム2012
- システム開発文書品質研究会（ASDoQ）
  - 藤田悠事務局長
  - ASDoQ大会2012
- 日本翻訳連盟、アジア太平洋機械翻訳協会

# 機械翻訳の活用に向けて

- 既存の資源の活用

- システム
- 技術



# 制限言語

- 簡易な制約を用いた文書作成
  - 日本語原文の読みやすさを悪化させない
  - 英語訳文の読みやすさを向上させる
  - ✓ 企業文書と自治体ホームページで検証
- 国際標準化の動き
  - ISO/AWI 24620-1
    - Language resource management
    - Simplified natural language
    - Part 1: Basic concepts and general principles

# 対訳用語集

- 高精度で翻訳するには辞書の整備が必要
  - 新規の語彙
  - 単語より長い、ひとまとまりの意味のある語句
- これらの自動抽出が必要



- 意味のある語句の自動抽出法
  - 接続関係を用いた用語(語句)抽出手法の利用

# 問題設定

- 最終的には人が見て「用語」を判定
  - ① 文書集合から用語候補を自動抽出
  - ② 用語候補を（正解を落とさないようにしつつ）できるだけ数を減らす（人手でチェックできる程度まで減らす）
  - ③ 人間が候補を見て、対訳用語集に含めるべき用語を決定

# 用語抽出実験

- 文書集合から目的用語を抽出
  - 文書集合:自動車メーカーの整備マニュアル
  - 英文103,828文
  - 取り出すべき用語:翻訳会社から提供を受けた用語集の語句(英語)100語句

# 実験結果

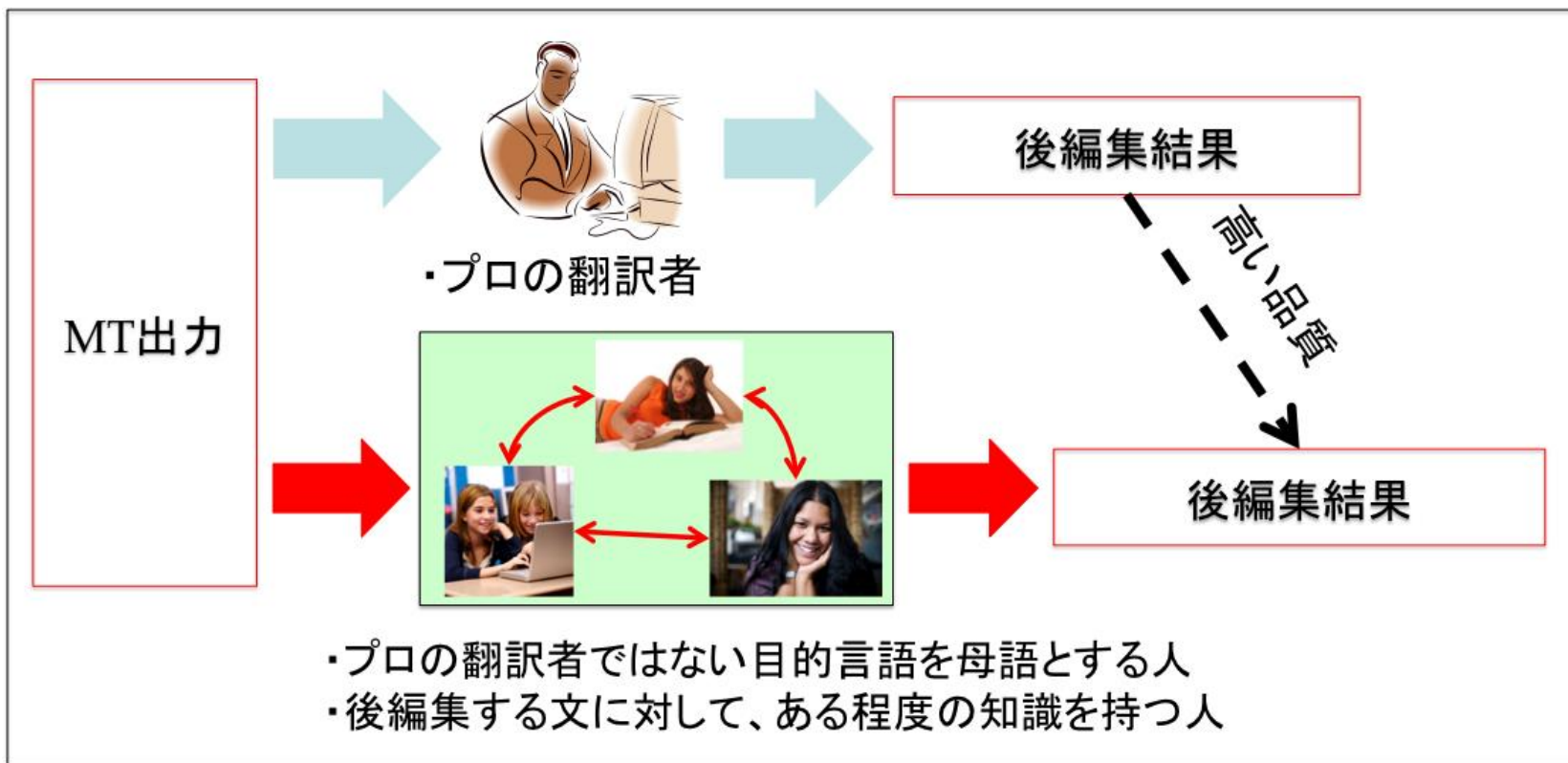
- 統計情報を用いて自動で専門用語候補を抽出
    - 65,712個（正解語句83個を含む）
    - 人手で見るのは困難
  - ヒューリスティックスで間引き
    - 4,947個（正解語句72個を含む）
    - 人手チェックが可能
- 人間による対訳用語集作成の支援に有効



# 後編集

- 目的
  - 後編集のコストを抑える
    - ボランティアベース
    - プロの介入を最小限に抑える
- 仮説
  - プロの翻訳者よりも、分野知識のある素人母語話者の方が、より良い後編集ができる
    - クラウドソーシング(集合知)後編集の提案

# クラウドソーシング（集合知）後編集



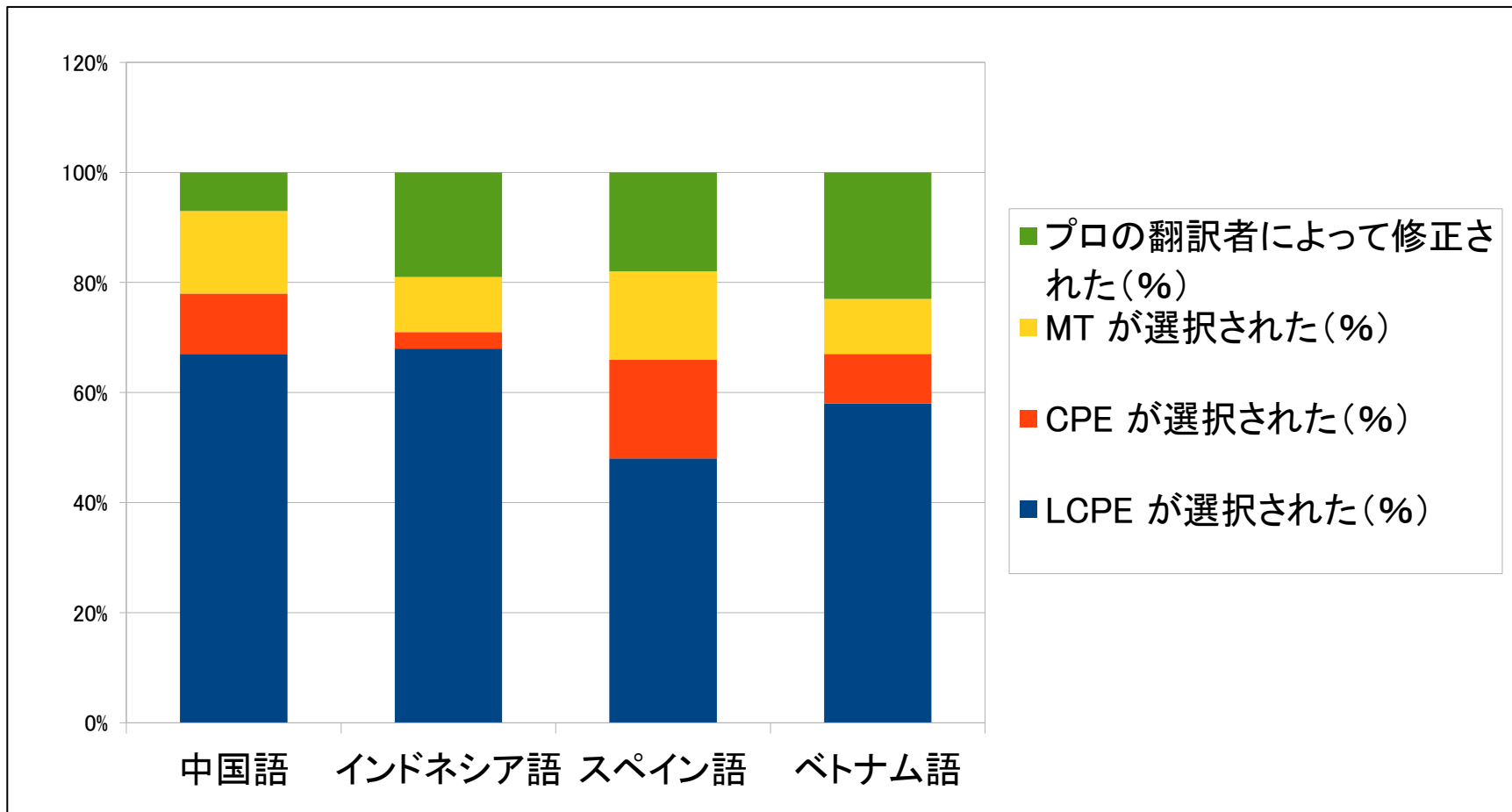
# 実証実験

- 豊橋技術科学大学の留学生に英語版ホームページの母語への機械翻訳出力の後編集を依頼
  - 対象言語
    - アラビア語、インドネシア語、ポルトガル語、スペイン語、中国語、ベトナム語、フランス語、ドイツ語、韓国語
  - 対象となるページ数：64ページ
  - 対象となる文数：約2,500文

# プロの翻訳者による評価

- プロの翻訳者に入力文に対しての機械翻訳結果と後編集結果をランダムな並びで提示
  - 入力文に対して相応しい訳文を選択
  - 候補の中に十分なものが無かった場合は修正

# 評価結果



**CPE**: 後編集された出力

**LCPE**: 最後に後編集された出力

# 考察

- 通常の後編集の場合
  - MT出力の84～90%は後編集が必要
- クラウドソーシング後編集の場合
  - 66～82% はMT出力か、LCPEが選択される
  - プロの後編集者は、MT出力とLCPEを見て、いずれにも満足できない場合(18～34%)に編集作業を行えばよい。

# TERによる自動評価

- MTよりLCPEを修正する方が作業量は少ない

言語	MT出力とLCPE間	LCPEとプロによる修正文間
中国語	54	27
インドネシア語	38	26
スペイン語	40	34
ベトナム語	54	27

# 考察

- 通常の後編集の場合
  - MT出力の84～90%は後編集が必要
- クラウドソーシング後編集の場合
  - 66～82% はMT出力か、LCPEが選択される
  - プロの後編集者は、MT出力とLCPEを見て、いずれにも満足できない場合(18～34%)に後編集を行えばよい。
  - LCPEを対象に後編集すれば、MT出力を対象にする半分の労力で済む



- 国際標準化の動き
  - ISO/TC 37/SC5 N 41  
Translation services
    - Requirements for machine translation (MT)  
and post edition levels
- TC 37/SC 5 Translation, interpreting and related technology



## TAUS EXECUTIVE FORUM 2013

- Calendar >
- Conferences >
- Executive Forums >
- Roundtables >
- Workshops >
- Online >
- Awards >
- Past Events >



*Hosted by Oracle in Tokyo (Japan) on April 11-12, 2013*

- Overview** Call for Proposals Program Venue Registration

TAUS Executive Forums are two-days meetings for buyers and providers of language services and technologies, aimed at an open exchange about language business innovation and translation technology. The TAUS Tokyo Executive Forum 2013 is the third TAUS meeting in Japan. The first took place in April 2010 and was also hosted by Oracle Japan.

### To Open or Not to Open

An orderly world, that's what we want. Knowing who owns what. Everything in its place. Clear boundaries.

But citizens are publishers en masse. They are our trusted sources. Globalization is enabling unprecedented cultural exchange. Non English speaking emerging markets and their new consumer classes are driving global growth. Information technology is advancing and increasingly embedded into our lives. These trends are converging to

# 終わりに

- 産業日本語(情報発信型の日本語文書)について、様々な立場からの議論が進むことを願っております。
- 国内外の活動の連携による新しい動きを期待しています。
- ご参加の皆様からのご意見を、ぜひお願いいたします。
- この発表の内容に興味を持たれた方はご連絡ください。
- データ(文書)を提供していただけると大変にありがとうございます。