資料2-4

バオバブにおける自然言語処理関連データ構築の取り組みと課題
次世代人工知能社会実装WGへの期待

株式会社バオバブ

グローバルにおける日本の機械翻訳

Coling 2016（**計算言語学分野におけるトップの国際会議**）**@大阪**

BAOBAB

グローバルにおける日本の機械翻訳

前回Coling 2014（計算言語学分野におけるトップの国際会議）@ダブリン

日本勢の存在感は希薄

BAOBAB

初めまして　バオバブです。

弊社は、

1.翻訳

- 機械翻訳エンジンの提供
- 学習データ（対訳）をクラウドで構築

2.画像収集・アノテーション・タグ付け

3.音声データ収集・書き起こし

4.他　機械翻訳評価・対訳シナリオ作成　等

を主に手掛けております。

コーパス（対訳）ベース機械翻訳エンジンの課題

一般的に、コーパス（対訳）ベース機械翻訳エンジンの翻訳精度は学習させる対訳（学習データ）の量と相関。

課題：

日本では人力翻訳の単価が高い故、学習データ費用負担が高く学習データの規模に限界

機械翻訳構築のプロセス（例：アパレル特化自動翻訳エンジン）

1.アパレルサイト アイテム説明全テキストデータ（日本語原文）を分析

BAOBAB
（旧「留学生ネットワーク@みんなの翻訳」）
運営：株式会社バオバブ

下訳にNICTの機械翻訳エンジン採用

2.翻訳サイト「BAOBAB」にて、安価・自動翻訳エンジンにマッチするように日本語→多言語に人力翻訳（対訳構築）

今夏大流行のマキシ丈ワンピース。この夏1枚は欲しいアイテムです。

↓

Maxi length dresses are one of the hottest items of this summer. One of the items you definitely need to buy this summer.

3.システムにてモデル化

例：翻訳モデル

今夏→this summer
大流行→the hottest items
マキシ丈ワンピース→Maxi length dresses
この夏→this summer
1枚は欲しいアイテムです→are one of the items you definitely need to buy.

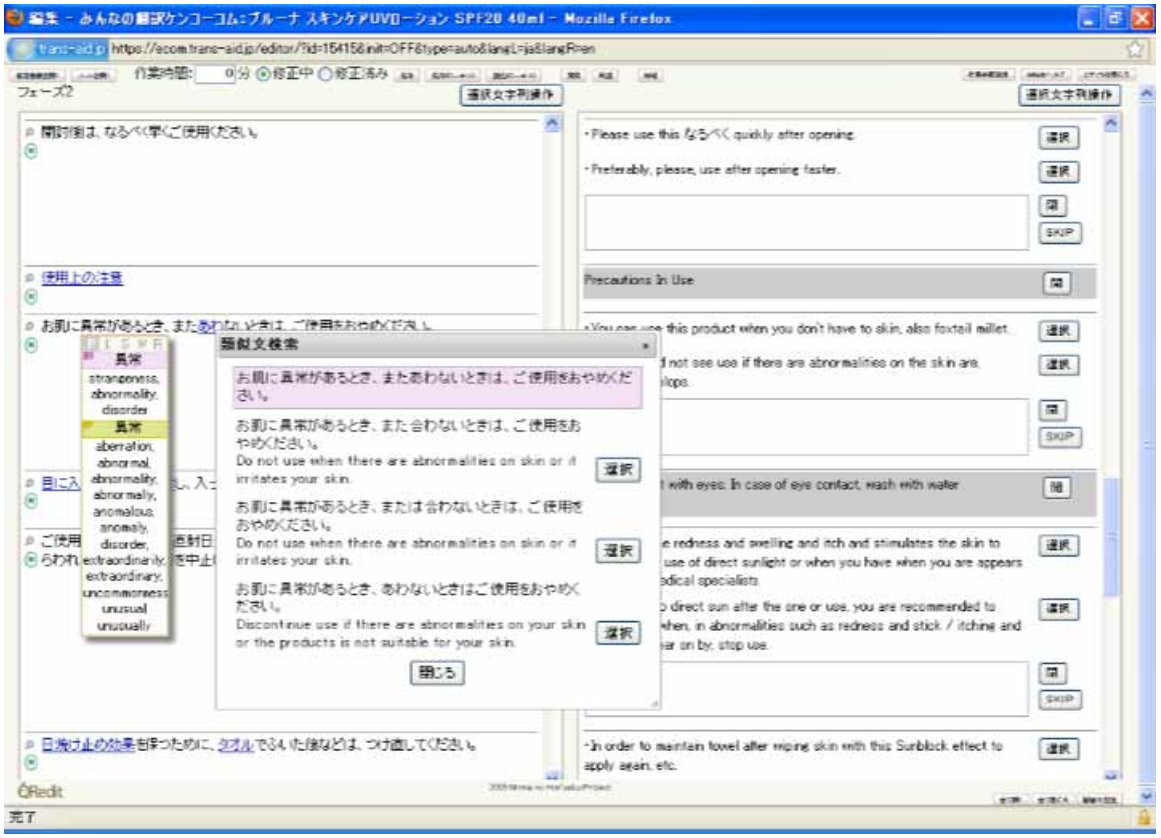
4.自動翻訳エンジン構築

他モデル
プログラム
言語モデル
（翻訳ルール）
翻訳辞書

機械学習・
テキストマイニング

BAOBABサイト（旧留学生ネットワーク@みんなの翻訳）

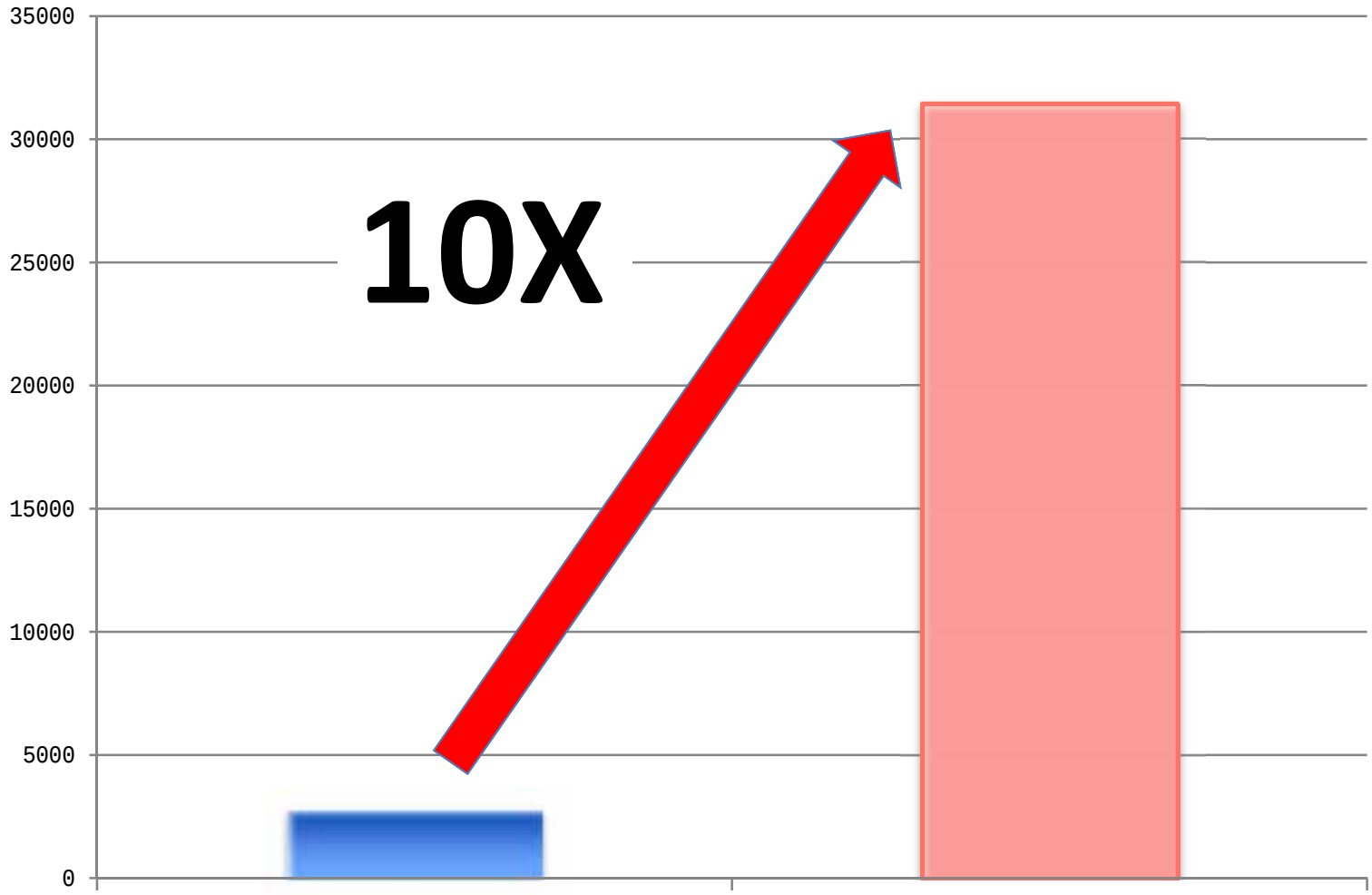
編集画面－高度翻訳支援による翻訳効率の向上



翻訳作業の下訳としてNICTの機械翻訳エンジンを実装することで、翻訳作業の生産性・効率向上を図る。

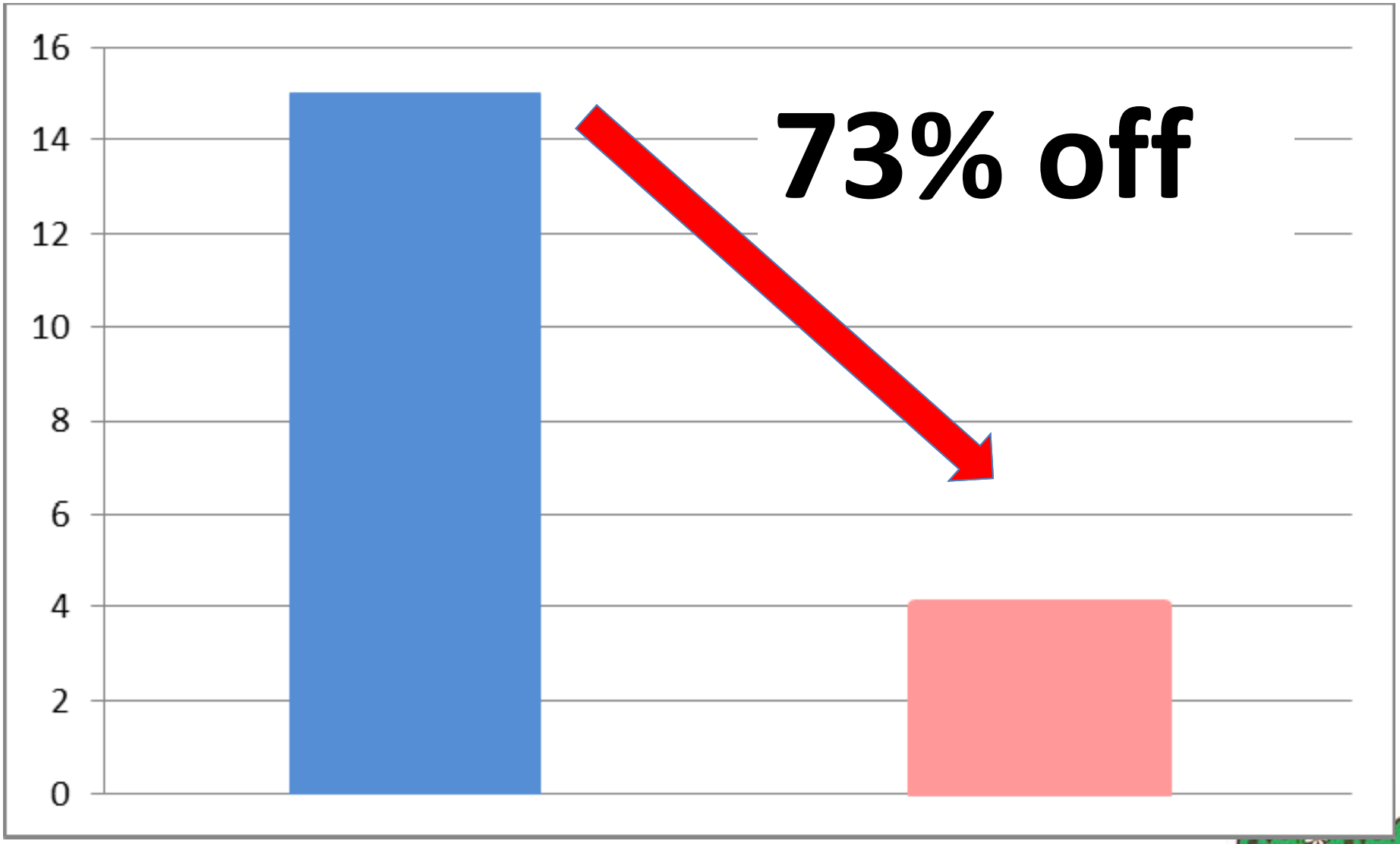
BAOBABサイト（旧留学生ネットワーク@みんなの翻訳）

1翻訳者あたりの翻訳文字数



BAOBABサイト（旧留学生ネットワーク@みんなの翻訳）

1文字あたりの翻訳単価



課題

私たちは切望しています。

課題

言語資源データと計算資源(GPU)

・日本語は話者数の多さの割に他言語対と比較して対訳データが少ない

・公開されているNTCIR (特許) やASPEC (科学論文) は分野に偏り

・要約・対話はそもそも大規模なデータがない

データがあるところに研究が生まれ民間企業での開発が進み実務化が普及する

課題

例：公共性高い固有名詞データ・一般分野対訳データ

宿名	住所（日本語）	住所（英語）	説明文
さがら亭	西臼杵郡五ヶ瀬町字五ヶ所字広木野145-68	145-68 Aza Hirokino, Gokasho, Gokase-cho, Nishiusuki-gun	果てしなく広がる海原と澄み渡るような静寂に包まれた森を楽しむ宿
ホテル　バオバブ	別府市鉄輪西3組	3-kumi Kannawanishi, Beppu-shi	スタッフ一同お客様との出会いを大切にいたします。別府での想い出はご家族の宝物になるでしょう。
瀞流の宿　中川	耶麻郡猪苗代町蚕養字沼尻山甲6543-145	6543-145 Numajiriyama-ko, Kogai, Inawashiro-machi, Yama-gun	家庭的な手作り料理と温かなおもてなしでごゆっくりとお寛ぎ頂けます。周囲には１年中、楽しめる観光地やシーズンスポットもございます。ロシア語対応可能です。

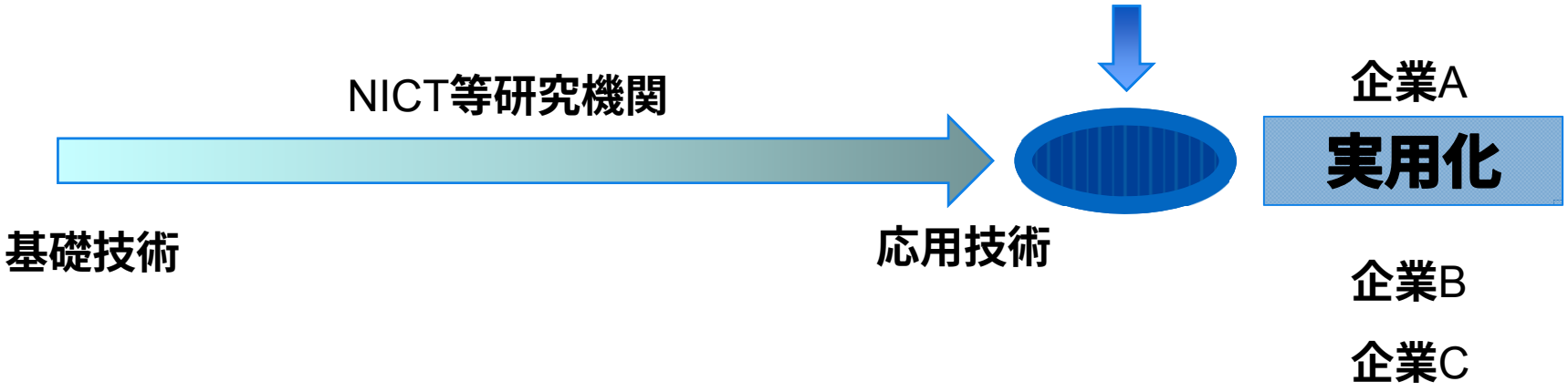
言語資源データ(何を)

具体的には（例）：

■種類
・対訳（社会実装転用可能な一般分野対訳・辞書）
・要約
・対話
■　分量
・100万文単位
■他
・きれい
・フリー
・誰でも目的を問わず簡単に入手出来る

言語資源データ(どのように)

悩ましい論点：



他　権利関係の整理　等

言語資源データ(まずは)

■出来るところから (現状の改善)

■少しでも (分野限定)

■早く (始めは生データでも可)

言語資源データ

付随（けれど重要）：

■法整備（権利関係）

例：ニューラルネットワークのような新手法であると、生成されたものに生成元のテキストを書いた人の権利がどこまであるか？

社会的議論が必要

課題

私たちは切望しています。

言語資源データの構築･拠出
及び法制度に関する社会的議論