言語横断テキストマイニングのための翻訳対抽出

那須川 哲哉　Daniel Andrade
海野 裕也　村松 祐希　山本 和英

発表者：06t4071f 林華

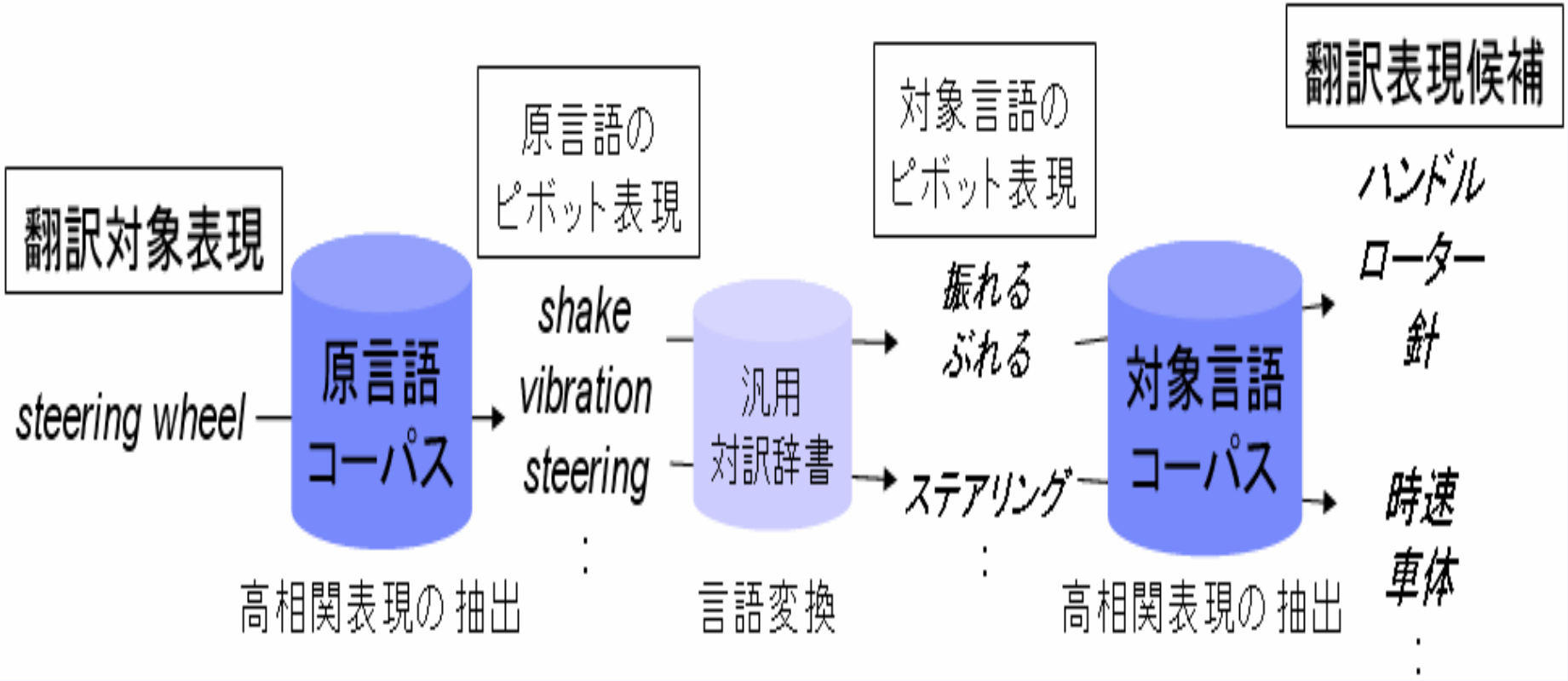
背景

- 企業活動のグローバル化により業務関連のテキストデータが多言語
- 現地法人の担当者のみがデータにアクセスしているのが現状である
- 多言語データを一元的に管理して網羅的に分析できれば、多様な地域で自社商品や競合商品への評価、不具合など把握できる

既存対訳辞書はダメ？

- 分析対象表現を他言語に直訳できない
  - ・例　　driver's side を「運転手側」ではなく
    「運転席側」に翻訳
- 量が膨大で、効率よくない

システム構築アプローチの概念図



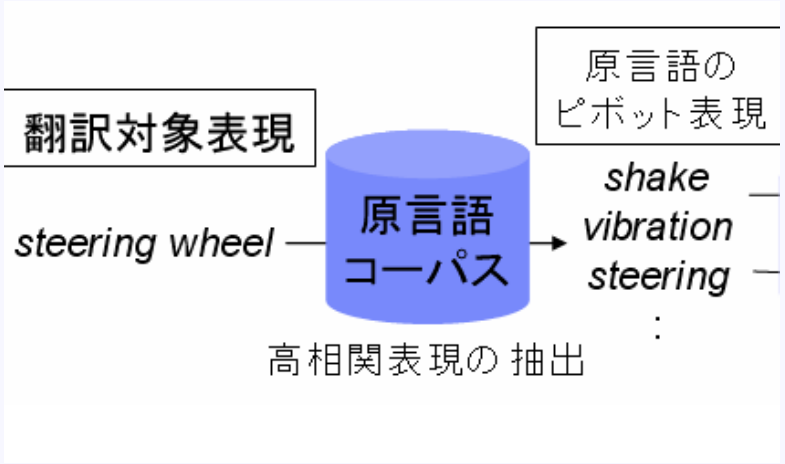
翻訳対抽出アルゴリズム

- 翻訳表現候補のリストを作成
- リスト中の各表現の翻訳表現としての妥当性を評価

翻訳表現候補リスト作成ステップ1

- 与えられた翻訳対象表現に対し、原言語コーパスにおいてその表現と相関の高い表現のリストを抽出

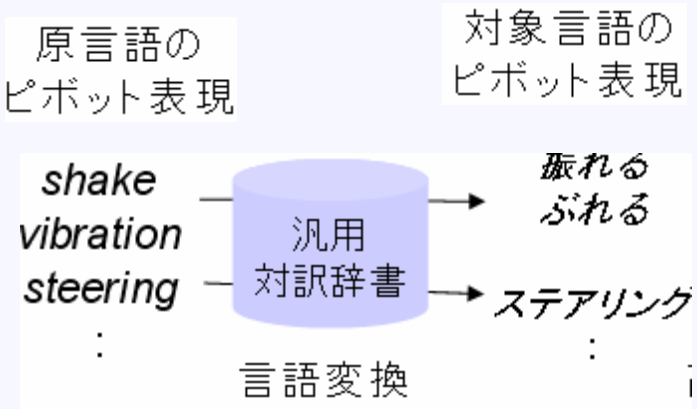
$$\frac{\dfrac{(\text{表現 A とB を両方含む文書数})}{(\text{表現A を含む文書数})}}{\dfrac{(\text{表現 B を含む文書数})}{(\text{全文書数})}}$$



- 説明：翻訳対象となる表現A を含む文書集合において表現B が出現する割合と、原言語コーパス全体において表現B が出現する割合との比を取ったものであり、その値が 1 を超えれば相関が強いということになる

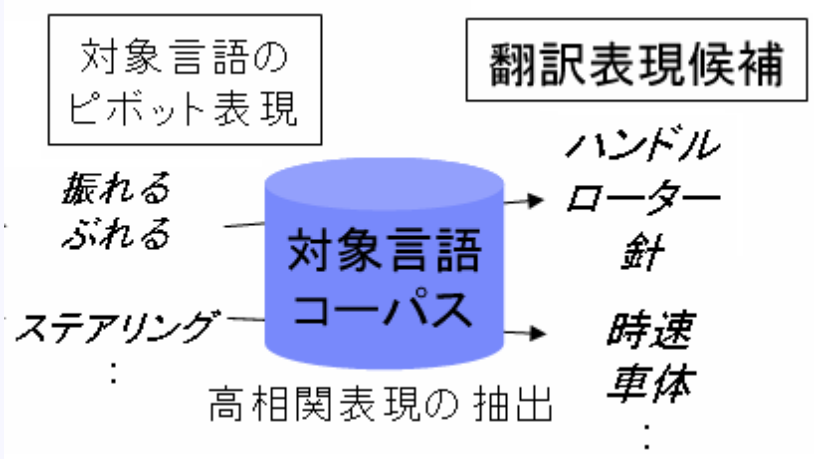
翻訳表現候補リスト作成ステップ2

汎用対訳辞書を用いて、第一段階で得られた相関の高い表現を対象言語に変換する。その際、汎用対訳辞書の見出しに含まれない表現は単純に対象外とし、訳語候補が複数存在する場合には、全ての候補を出力する



翻訳表現候補リスト作成ステップ3

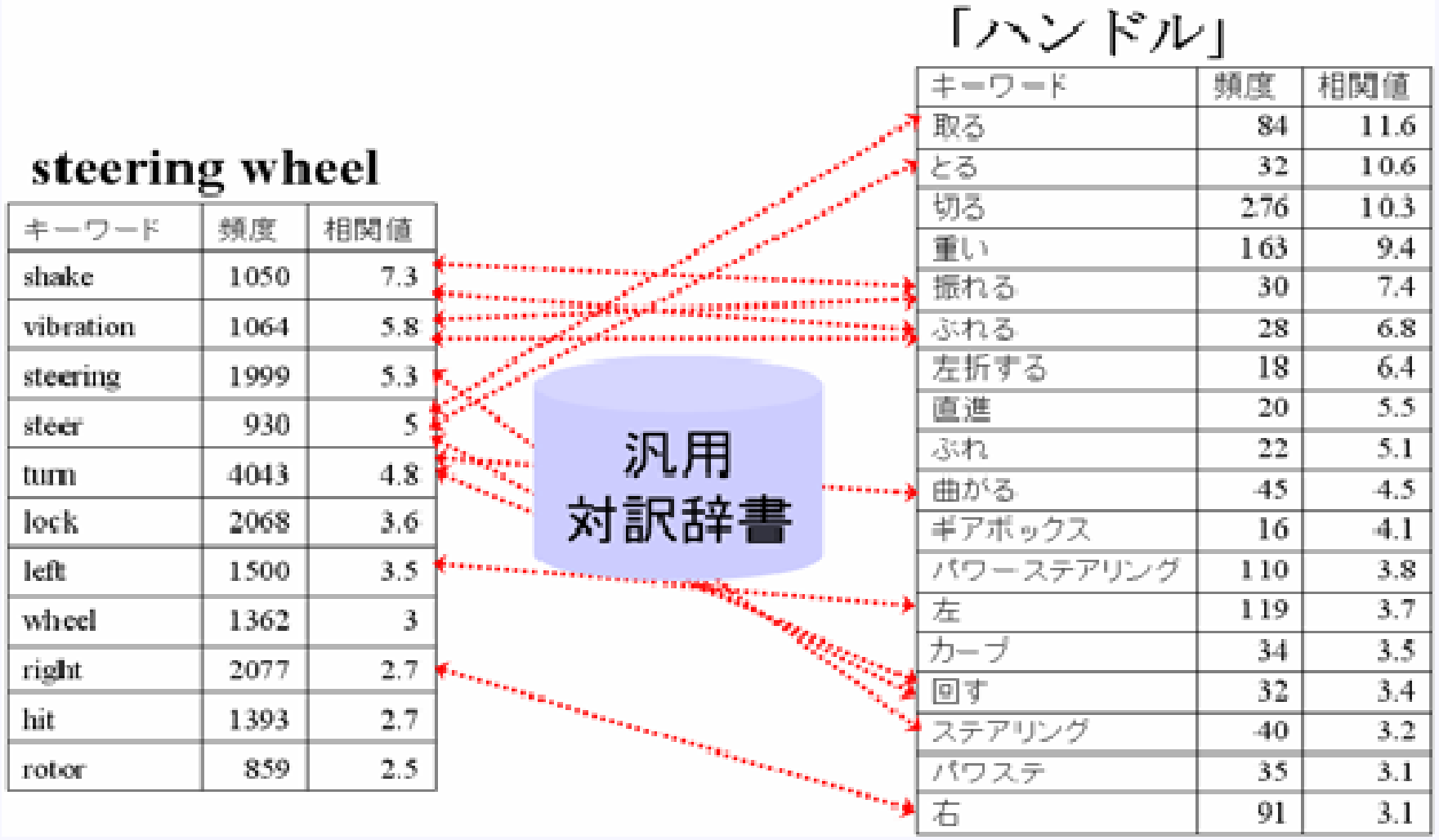
　第三段階では、第二段階で得られた各表現に対し、第一段階と同様に対象言語コーパス中で相関の高い表現のリストを抽出する。その結果得られる各リストの表現をマージしたリストが、翻訳表現候補リストとして出力される



翻訳表現としての妥当性評価

前ステップで得られた各翻訳表現候補と翻訳対象表現との意味的類似性を測る尺度として、各表現に相関の高い語が汎用対訳辞書を介してどれだけ対応しているかを調べる

例：steering wheel と「ハンドル」の意味的類似性の評価



実験データと実験環境

- 国土交通省自動車交通技術安全部審査課の「自動車不具合情報」 20,269 件
- 米国政府組織に属するNational Highway Traffic Safety Administration (NHTSA)の “Consumer Complaints2” 525,055 件
- 翻訳対象表現が辞書登録ずみ表現のうち部位のカテゴリに属する表現を用いた。英語は100表現、日本語は30 表現である。
- 汎用対訳辞書としては英日機械翻訳辞書
- E→J→E→J という展開を行った

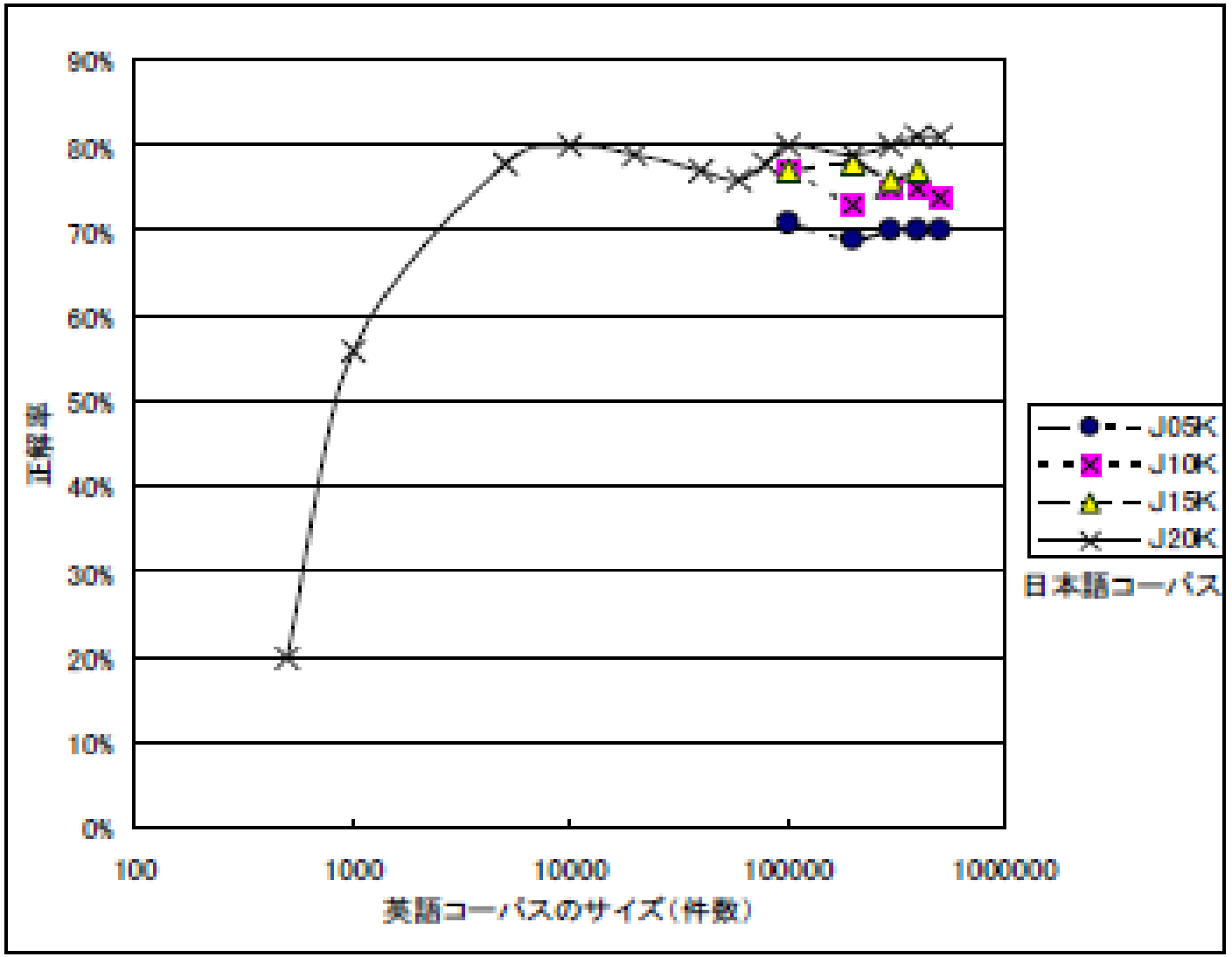
翻訳対抽出精度

	英語→日本語 (100 表現)		日本語→英語 (30 表現)	
	件数	精度	件数	精度
1 位の候補が正解	31	31%	10	33%
5 位以内に正解	60	60%	20	66%
10 位以内に正解	72	72%	22	73%
20 位以内に正解	81	81%	27	90%

翻訳対抽出の出力例
（下線の表現を正解と判断）

翻訳対象	radiator	headlight	ガソリン	ブレーキ
第 1 候補	冷却	前照灯	gasoline	foot
第 2 候補	冷却水	ヘッドライト	smell	pedal
第 3 候補	ラジエータ	ライト	fuel	parking
第 4 候補	ホース	レンズ	gas	brake pedal
第 5 候補	タンク	方向指示器	tank	lane
第 6 候補	ラジエター	熱	fuel tank	braking
第 7 候補	オーバーヒート	水滴	odor	park
第 8 候補	コア	曇り	gas tank	brake
第 9 候補	ヒーター	制動灯	exhaust	traffic
第 10 候補	水漏れ	雨天	leak	someone

コーパスのサイズと精度の関係



分析の結果

サブカテゴリ／キーワード	brake 63302	engine 47561	tire 40179	transmission 34307	light 30825	seat 22596	gear 19719	door 18925
Model A 15124	1309 0.7	881 0.6	3761 3.1	1166 1.1	692 0.7	503 0.7	656 1.0	640 1.1
Model B 13574	1051 0.6	1461 1.1	1147 1.0	1402 1.5	560 0.6	205 0.3	540 0.9	237 0.4
Model C 10058	1029 0.8	824 0.8	1061 1.3	346 0.5	320 0.5	220 0.4	328 0.8	485 1.2
Model D 9405	987 0.8	498 0.5	233 0.3	958 1.4	660 1.1	422 0.9	379 0.9	581 1.5
Model E 9167	1710 1.5	739 0.8	282 0.3	627 0.9	312 0.5	295 0.6	620 1.6	813 2.3
Model F 8808	935 0.8	1343 1.6	415 0.5	1373 2.2	700 1.2	187 0.4	523 1.4	546 1.5

続き―ある表現tankの相関値が高い場合

	キーワード	頻度		相関値	
	fire hazard	20		11.7	
	leakage	29		6.9	
	fuel leak	9		3.8	
	leaking	16		2.9	
	fire	74		1.9	
	malfunction	14		1.3	

おわりに

実装実験では、英語から日本語へ、日本語から英語へ、同等以上の精度を出すことができ、本手法の汎用性の高さを確認できた。