Memsource機械翻訳レポート

2021年第四半期

MEMSOURCE

Memsourceの機械翻訳レポートとは？

- 四半期ごとに機械翻訳の調査を実施
- お客様が実際のビジネスにおいて、Memsourceのシステムで行った翻訳のデータを使用

調査方法について

ポストエディットデータのスコアづけ

- ポストエディット（PE）：機械翻訳で出力された文章を、翻訳者が修正・改善し、より質の高い翻訳文を完成させるプロセス

- スコアづけ：
  機械翻訳の出力がそのまま受け入れられた場合：
  スコア＝100
  翻訳者によって完全に書き換えられた場合：
  スコア＝0→スコアが高いほど、機械翻訳の質が高いことを示唆

調査方法について

コンテンツの分野別の分析

- コンテンツの分野（＝ドメイン）によって、機械翻訳の出力品質が大きく変わる
- Memsource Translateでは、文中のキーワードをAIが分析し、自動的にドメインを分類

DOMAIN	KEYWORDS
1. Medical	‘study’, ‘patients’, ‘patient’, ‘treatment’, ‘dose’, ‘mg’, ‘clinical’
2. Travel and Hospitality	‘km’, ‘hotel’, ‘guests’, ‘room’, ‘accommodation’
3. Business and Education	‘team’, ‘business’, ‘work’, ‘school’, ‘students’
4. Legal and Finance	‘agreement’, ‘company’, ‘contract’, ‘services’, ‘financial’
5. Software User Documentation	‘click’, ‘select’, ‘datat’, ‘text’, ‘view’, ‘file’
6. Consumer Electronics	‘power’, ‘battery’, ‘switch’, ‘sensor’, ‘usb’
7. User Support	‘please’, ‘email’, ‘account’, ‘domain’, ‘contact’
8. Cloud Services	‘network’, ‘server’, ‘database’, ‘sql’, ‘data’
9. Industrial	‘mm’, ‘pressure’, ‘valve’, ‘machine’, ‘oil’
10. Software Development	‘value’, ‘class’, ‘type’, ‘element’, ‘string’
11. Entertainment	‘game’, ‘like’, ‘get’, ‘love’, ‘play’, ‘go’

表：ドメインと関連キーワード

調査方法について

使用したデータ

- 2021年1月～6月に収集したデータを使用
- 対象：
  機械翻訳の出力を翻訳者がポストエディットしたセグメント機械翻訳を利用できる状況で、翻訳者が一から翻訳を行ったセグメント
- 翻訳メモリのマッチングによる翻訳や、翻訳不能なセグメントは除外

結果

言語ペア別のスコア

- 多くの言語ペアで70点以上を獲得
- スペイン語、イタリア語、ポルトガル語など関連性の高い言語ペアで高スコア
- 形態が複雑なスラブ系言語ではスコアが下がる傾向

SOURCE LANGUAGE \ TARGET LANGUAGE	Czech	German	English	Spanish	French	Italian	Japanese	Portuguese	Russian	Chinese (Simplified)
Czech		87	76			76				
German			80	75	70	72				
English	67	71		77	73	73	66	75	67	68
Spanish		70	81		77	91		83		
French		63	74	70		74				
Italian		73	75	82	79			80		
Japanese			68							52
Portuguese			77	79						
Russian	71		77			64				
Chinese (Simplified)			57				48			

結果

前回調査からの変化

前回調査からスコアが上昇したのは22言語、低下したのは4言語

- 全言語ペアの平均スコアは71から73へと上昇
-

SOURCE LANGUAGE \ TARGET LANGUAGE	Czech	German	English	Spanish	French	Italian	Japanese	Portuguese	Russian	Chinese (Simplified)
Czech		4	-1			4				
German			1	-1	-2	-2				
English	2	1		0	0	1	0	1	1	1
Spanish		-4	1		0	N/A		0		
French		1	2	0		0				
Italian		1	1	4	3			-1		
Japanese			1							1
Portuguese			0	1						
Russian	N/A		2			N/A				
Chinese (Simplified)			3				4			

結果

人気の高い言語ペア

- 英日翻訳が2位、日英翻訳が4位と、日本語が関係するペアが上位に

- トップ10の言語ペアがMemsourceの全機械翻訳の54%を占める

1位	英語 ー スペイン語
2位	英語 ー 日本語
3位	英語 ー フランス語
4位	日本語 ー 英語
5位	英語 ーロシア語
6位	英語 ードイツ語
7位	英語 ーポルトガル語
8位	英語 ー中国語
9位	英語 ーイタリア語
10位	オランダ語 ー英語

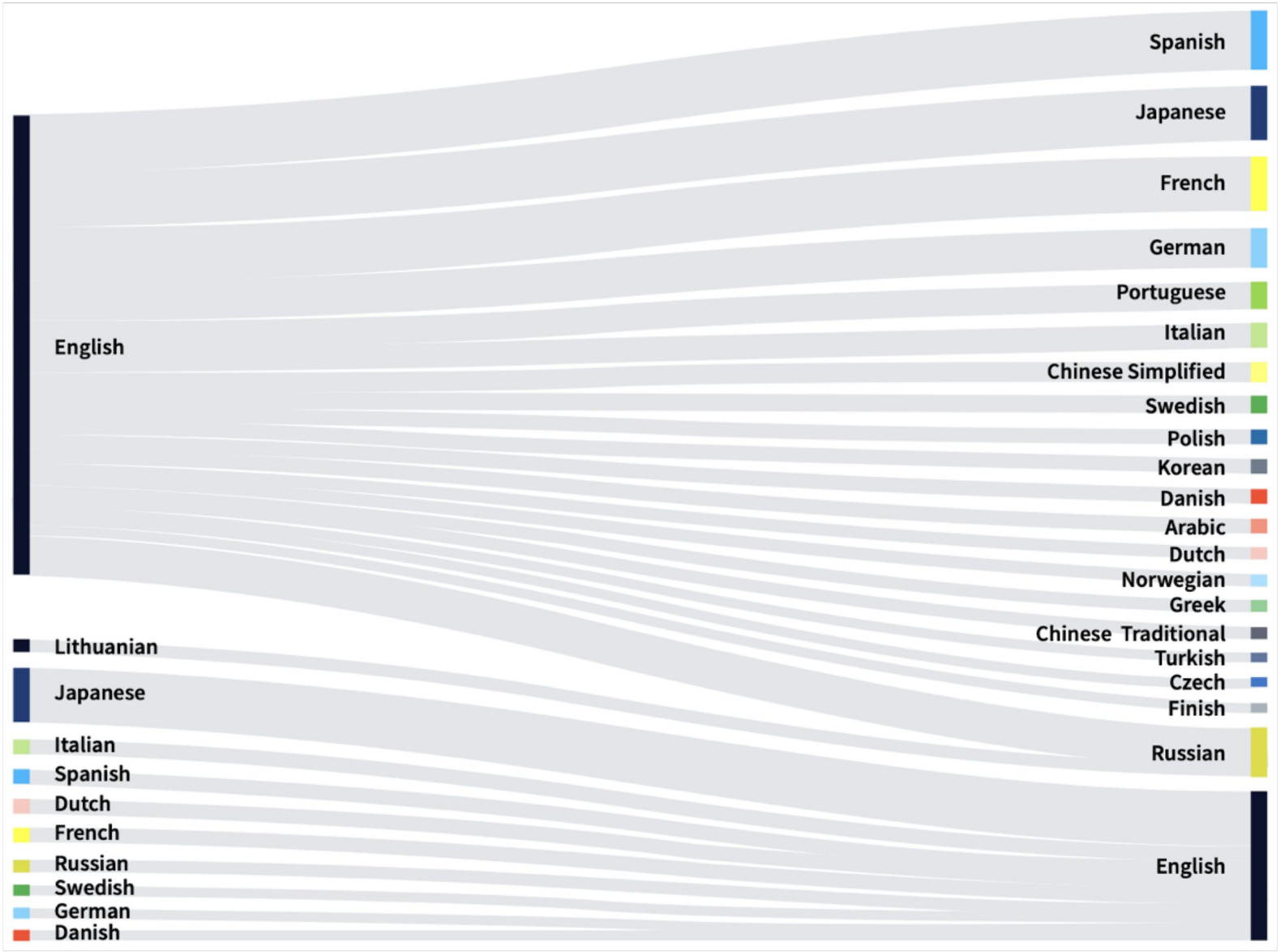
表：利用されている言語ペアのトップ10

結果

人気の高い言語ペア

- 英語をソース言語とする翻訳が多数を占める

- 英語への翻訳では日本語がトップ



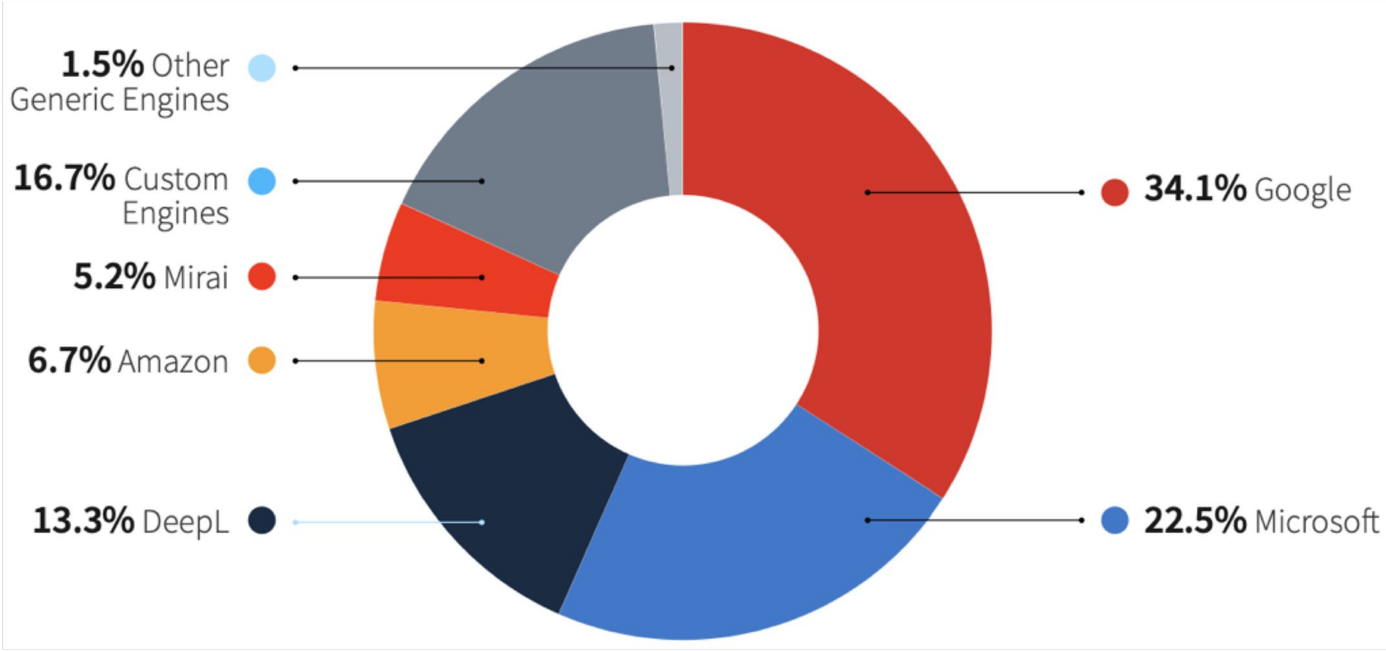
図：どの言語からどの言語へ翻訳されたか

結果

機械翻訳エンジン別の分析

- Memsourceでは30以上の機械翻訳エンジンをサポート

- Google、Microsoft、DeepL、Amazonなどの汎用エンジンが使用の大半を占める



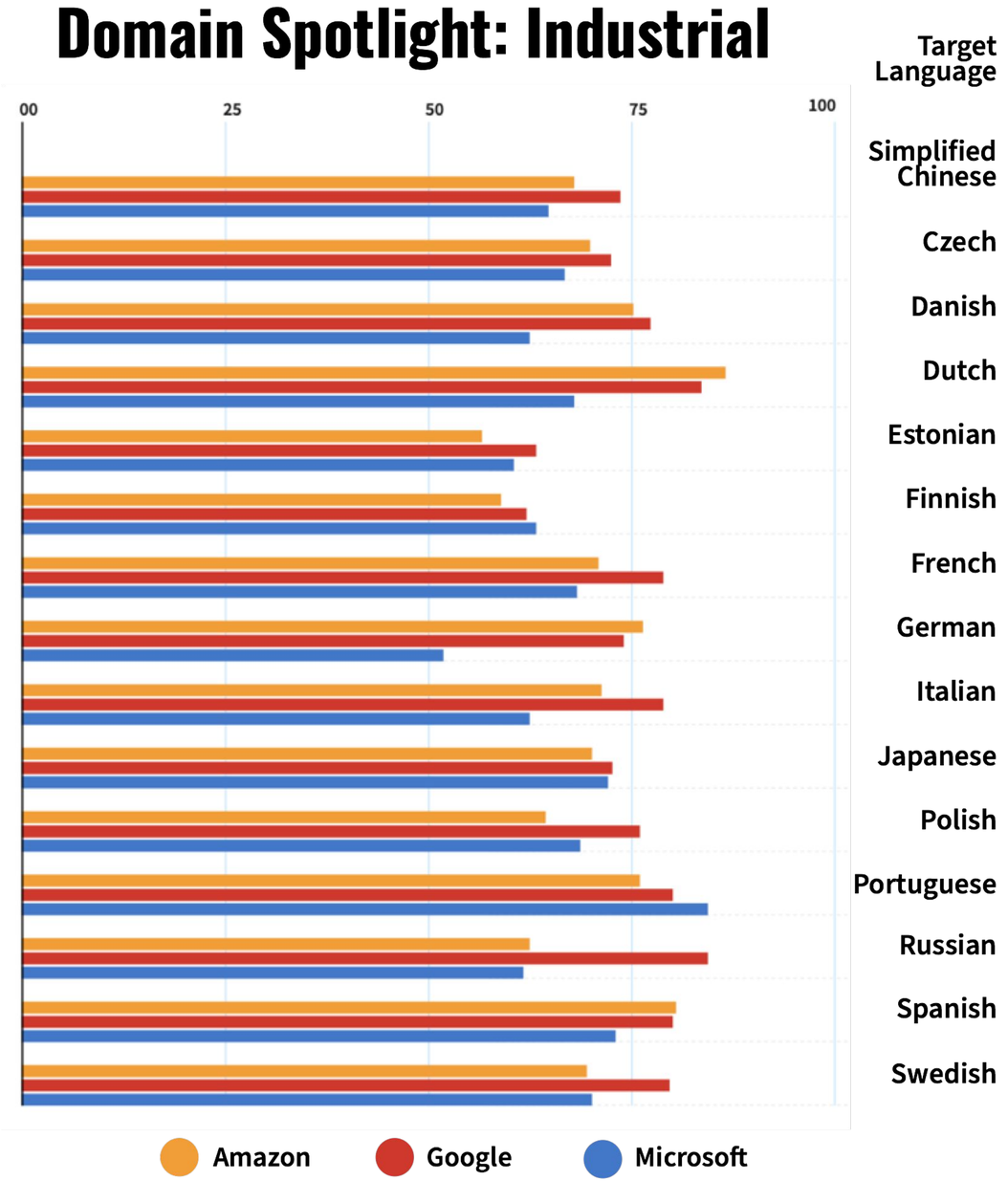
図：使用された機械翻訳エンジンの割合

結果

ドメインと機械翻訳エンジンの分析

- 今回の調査では「産業」ドメインに着目

- 機械翻訳のパフォーマンスは、言語とエンジンの両方によって大きく変化する



図：産業ドメインにおける、ターゲット言語別の機械翻訳エンジンの平均スコア

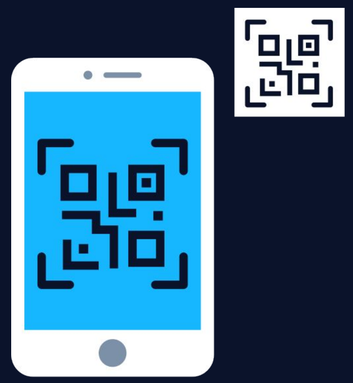
まとめ

- 機械翻訳エンジンのパフォーマンスは、言語ペアやコンテンツの種類によって異なる
- 機械翻訳の出力品質は全般的には高いと言える
- 特定の言語ペアやドメインに対して、どのエンジンを選択するかによって、出力品質が大きく変わってくる

より詳しく知りたいですか？

第四半期に公開された全文レポート（英語）の送付をご希望の方は
下記のメールアドレスまでお問い合わせください。

japan@memsource.com



MEMSOURCE