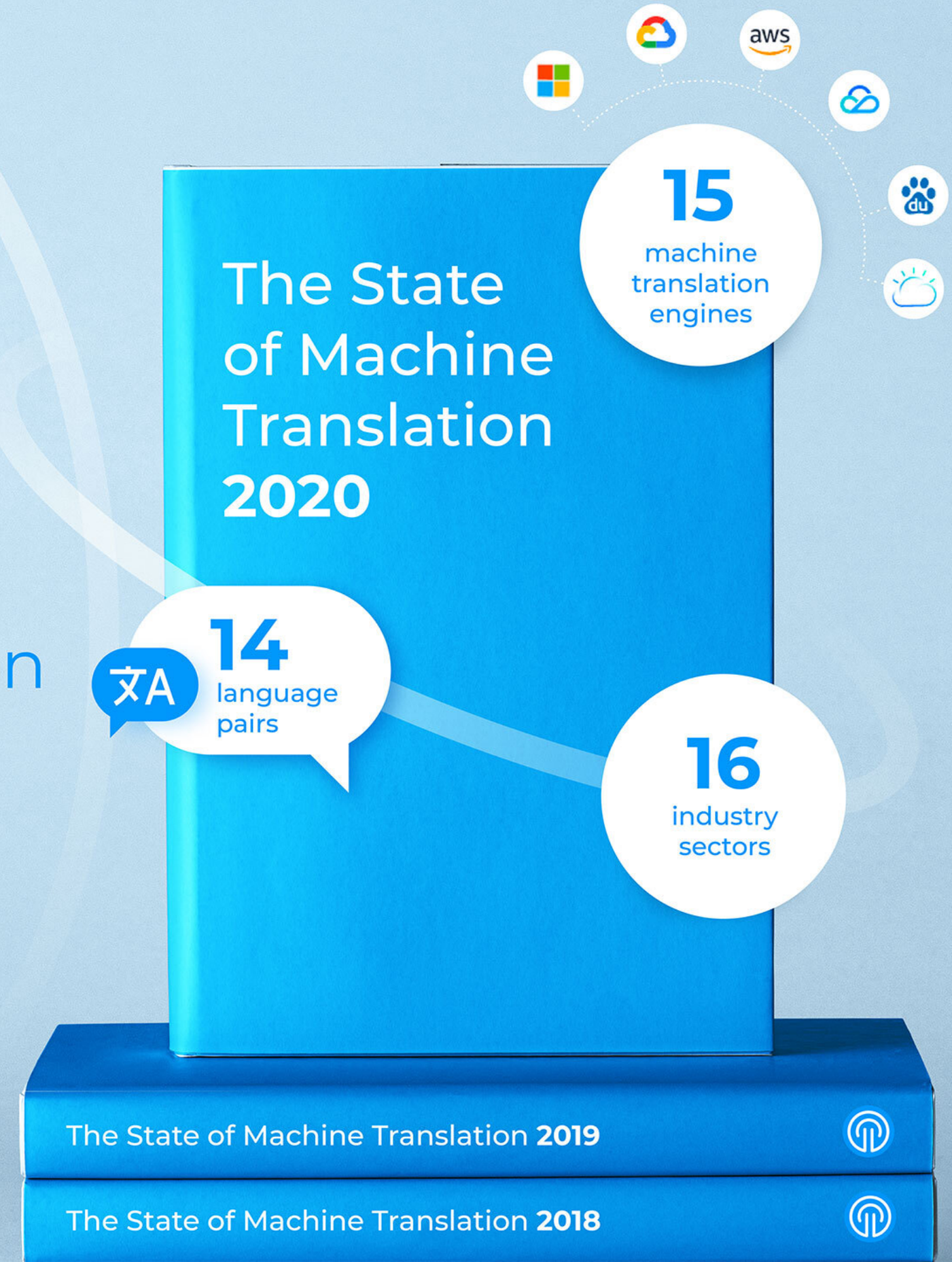
Independent multi-domain evaluation of commercial machine translation engines

In partnership with TAUS
the language data network

https://inten.to



DISCLAIMER

The MT systems used in this report were accessed from March 15 to April 5, 2020. They may have changed many times since then.

—

This report demonstrates the performance of those systems exclusively on the datasets used for this report (see slide 9) using proximity scores. The final MT decision requires Human LQA and depends on the use-case.

—

The evaluation is done on plain text data. We often see different results for tagged text (one you have in CAT/TMS) for some TM vendors and language pairs due to imperfect inline tag support.

—

The data originates from several large companies and also it's available for for purchase via TAUS Data Cloud. Some MT providers could have access to such data in the past and could have used it for training their models.

—

We run multiple evaluations for our clients for various language pairs and domains, observing different rankings of the MT systems than provided in this report.

—

There's no "best" MT system. Performance depends on how your data is similar to what they used to train their models and on their algorithms.

Executive Summary

The scores are dead, long live the scores! New semantic similarity scores (e.g. BERTscore) solve the main issue of syntactic similarity score (e.g. BLEU) - dealing with alternative translations and synonyms.

Each of 15 MT engines is best at something. 7 of them are enough to get the best quality for all 15 industries and 14 language pairs we have tested. For any given industry, 1-4 engines are enough for these languages.

The highest MT quality is available for Computer Software, Legal Services, and Telecommunications, with Software Strings and Documentation, Support Content, Policies, Processes and Procedures being the easiest content types.

Everything in Professional and Business Services, as well as Instructions for Use and Sales & Marketing Content in other industries are the hardest nuts for MT to crack.

Incredible spike in language coverage: +2,000 language pairs since June 2020, many low-resource languages added (see slide 33).

MT Landscape continues to evolve: 11 more vendors offer pre-trained MT engines since the June 2019.

About

We have been evaluating models for Machine Translation since May 2017 (Custom NMT as well).

—

As we show in this report, the Machine Translation landscape is complex, with models from 8 different vendors required to get the best quality across popular language pairs and a 90x difference in price. And it changes often!

—

To evaluate on your own dataset, reach us at hello@inten.to

—

To conveniently use the best-fit MT across multiple enterprise scenarios, check out our Enterprise MT Hub (next slide).

Intento Enterprise MT Hub

One place to evaluate and manage MT

Universal API to all MT engines

Single MT dashboard

Smart Routing with retries and failovers

Connects to many CAT, TMS and CMS

Works with files of any size

Get your API key at inten.to

MAY BE DEPLOYED ON PRIVATE CLOUD

Overview

1 DATASETS

2 EVALUATION APPROACH

3 EVALUATION RESULTS

4 MISCELANEOUS

5 CONCLUSIONS

CALL FOR LSP VOLUNTEERS!

15
Machine Translation Engines

14
Language Pairs

16
Industry Sectors

8
Content Types

Machine Translation Landscape

GENERIC STOCK MODELS

AISA, Alibaba, Amazon, Baidu, DeepL, eBay, Globalese, Google, GTCom, IBM, iFlyTec, Kakao, Microsoft, Mirai, ModernMT, Niutrans, Naver, NICT, Omniscien, Prompsit, PROMT, Rozetta, SAP, SDL, Sogou, Systran, Tencent, Tilde, Yandex, Youdao

VERTICAL STOCK MODELS

Alibaba, Baidu, Cloud Translation, Iconic, Microsoft, Omniscien, PROMT, SAP, Systran

CUSTOM TERMINOLOGY SUPPORT

Amazon, Baidu, Google, IBM, Iconic, Microsoft, Rozetta, SDL, Systran, Yandex

AUTO DOMAIN ADAPTATION

Globalese, Google, IBM, Kantan, Microsoft, ModernMT, Omniscien, SDL, Systran

MANUAL DOMAIN ADAPTATION

Alibaba, Baidu, Cloud Translate, Iconic, Omniscien, PangeaMT, Prompsit, PROMT, SDL, Systran, Tilde, Yandex

Machine Translation Engines*

Evaluated in this Study

Customization options:

- ☐ - none
- ◹ - TM (auto)
- ◹ - glossary
- ■ - both

Alibaba Cloud eCommerce MT ☐	Alibaba Cloud General MT ☐	Amazon Translate ◹
Baidu Translate API ◹	DeepL API ◹	Google Cloud Translation API ■
GTCom YeeCloud MT ☐	IBM Watson Language Translator ■	Microsoft Translator Text API v3 ■
ModernMT Realtime ◹	PROMT Cloud API ☐	SDL BeGlobal ■
SYSTRAN PNMT ■	Tencent Cloud TMT API ☐	Yandex Translate API ◹

* We have evaluated general purpose Cloud Machine Translation services with pre-trained translation models, provided via API. Some vendors also provide web-based, on-premise or custom MT engines, which may differ on all aspects from what we've evaluated.

1 Datasets

Datasets - Origin

The datasets are provided by TAUS

Every element has the following attributes:

- source text
- reference human translation
- language pair
- industry sector
- content type

Data samples to reproduce the study are available from TAUS <david@taus.net> and Intento <hello@inten.to>

Datasets - Cleaning

The first pass of cleaning performed by TAUS

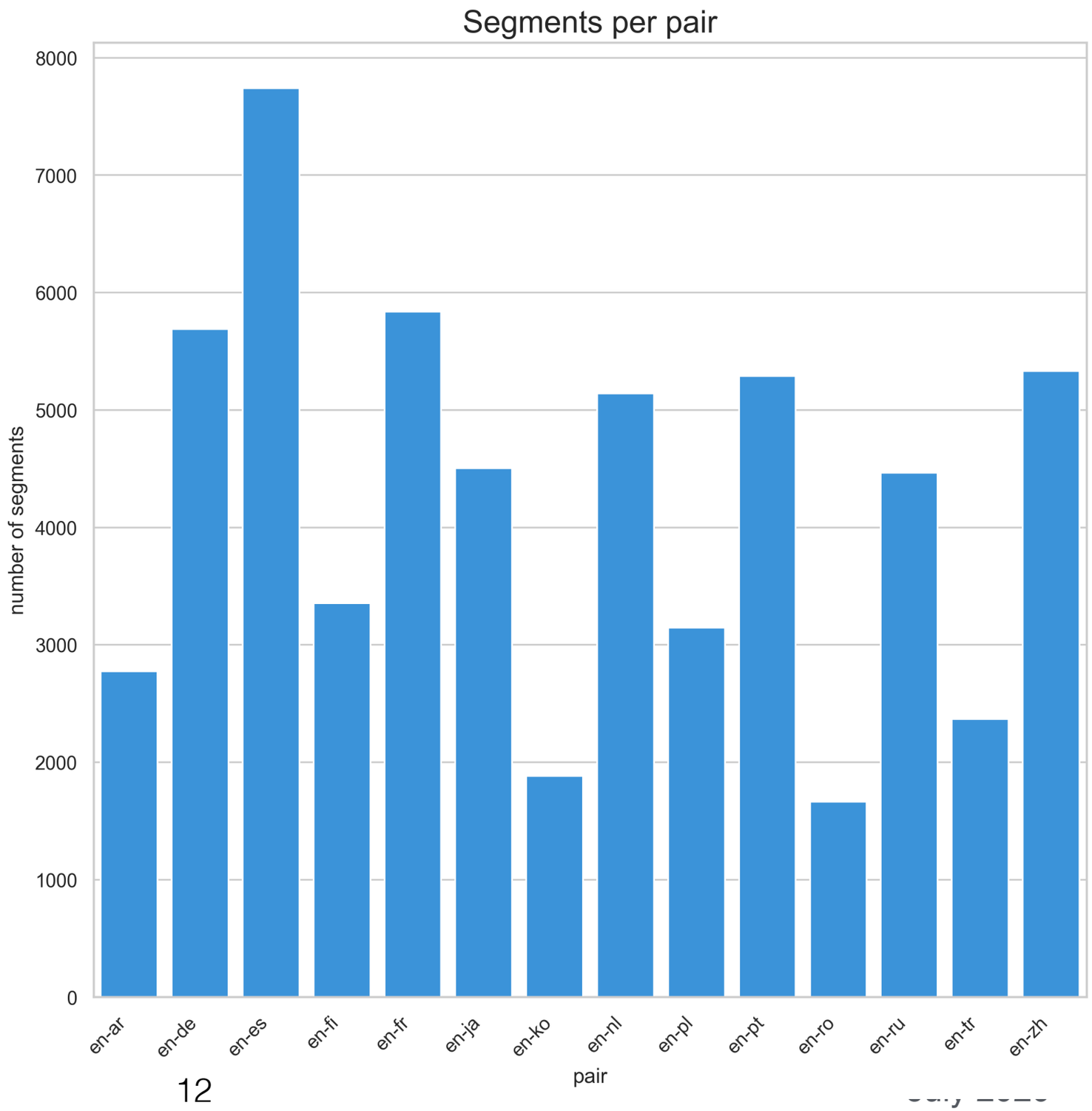
—

Additional cleaning by Intento:

- removed multi-sentence segments
- removed sentence fragments
- removed mistranslations using Automated Translation Quality Estimation (based on multilingual embeddings)
- removed out-of-domain samples using Automated Domain Clusterization (based on multilingual embeddings)

Datasets - Language pairs

14 language pairs, selected based on the availability of ~500 segments for several industry sectors.

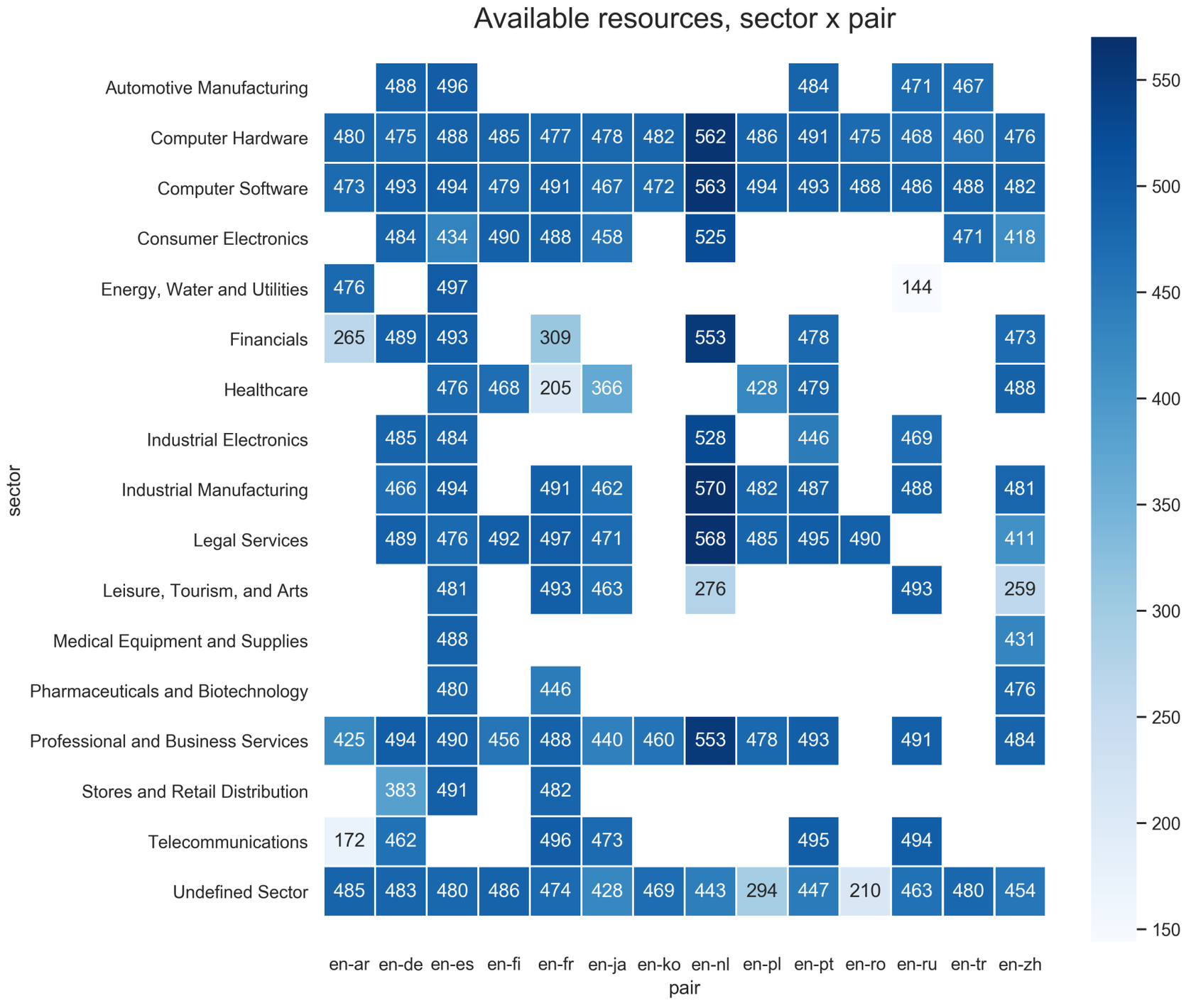


Datasets - Industry Sectors

16 industry sectors (+ Undefined)

4-15 industry sector per language pair

~ 500 segments per language pair per industry sector

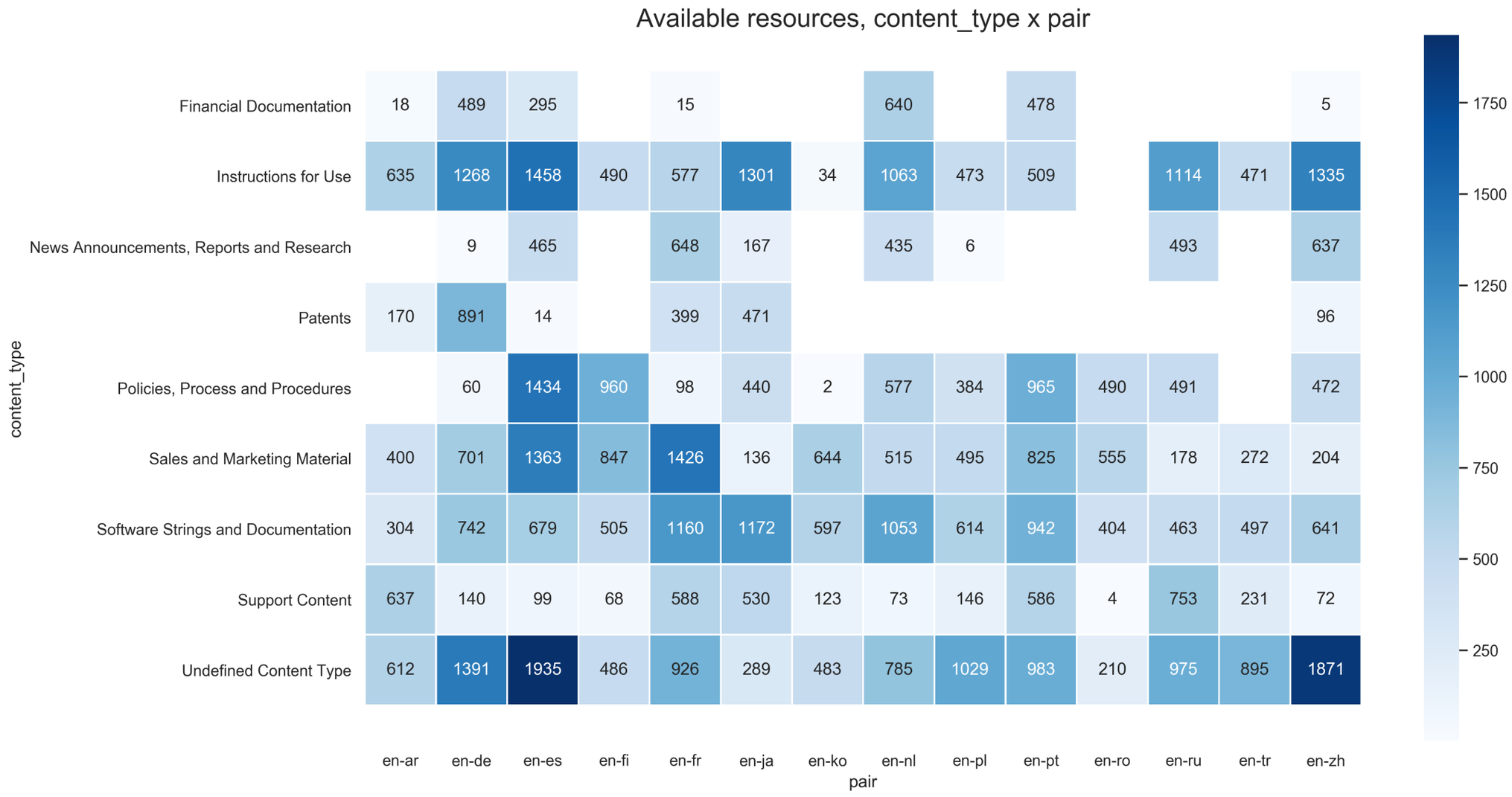


Content samples: Industry Sectors

Automotive Manufacturing	The liner is constructed of water-resistant nylon, and features an embroidered Bar & Shield logo.
Computer Hardware	Due to the best effort delivery of UBR traffic, it is typically the least expensive service offered by commercial carriers.
Computer Software	From the CT-KIP Activation Code drop-down list, select an activation code for the software token.
Consumer Electronics	Take appropriate safety measures against strong winds and earth-quakes to prevent the unit from falling.
Energy, Water and Utilities	Employer or employees shall not remove or deface labels on incoming containers of hazardous chemicals.
Financials	Inelastic demand — Total demand for a product that is not much affected by price changes, especially in the short run.
Healthcare	Substances known to cause increased methaemoglobin levels should thus be used with caution during therapy with inhaled nitric oxide .
Industrial Electronics	A continuous recorded section of approximately 5 seconds is required before the recording start point.
Industrial Manufacturing	Open system to preferred chemistry and not locked in to factory calibrated dyes.
Legal Services	Among those principles and purposes of the organization is the promotion of respect for human rights.
Leisure, Tourism, and Arts	You had three French fries and a couple bites of yogurt.
Medical Equipment and Supplies	When inhaling, top of yellow air float should be raised between blue arrows.
Pharmaceuticals and Biotechnology	N-oxides are oxidised forms of tertiary amines or oxidised forms of nitrogen containing heteroaromatic compounds
Professional and Business Services	Large legal firms often have multiple regional offices and can have branch offices distributed around the globe.
Stores and Retail Distribution	Add a dramatic touch to doorways at your celebration with these sparkly Hanging Glitz Black 50th Birthday Decorations.
Telecommunications	The following are the various commands referenced in this document:

Datasets - Content Types

8 content types (+ Undefined), uneven distribution across language pairs



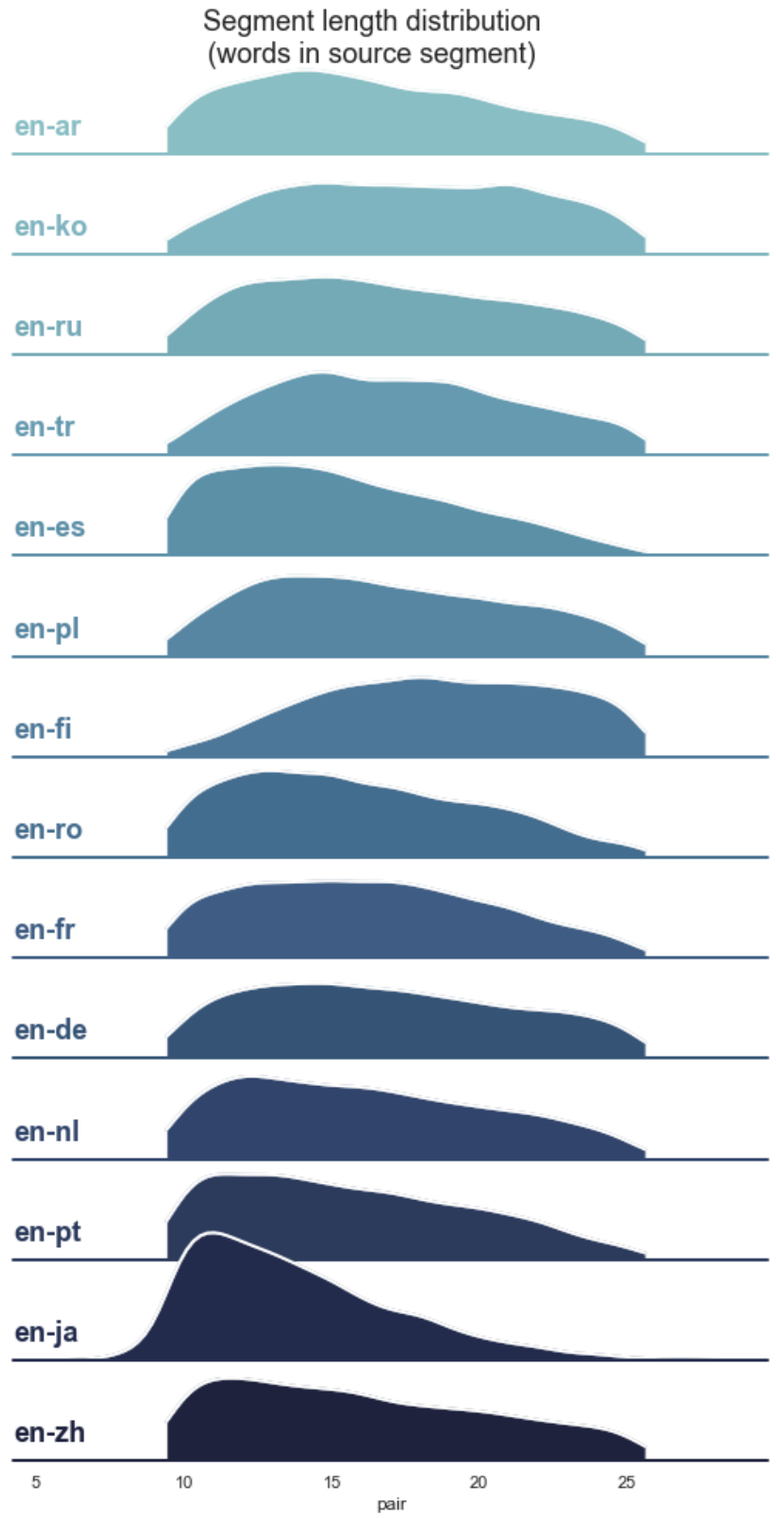
Content samples: Industry Sectors

Financial Documentation	The share shall be indivisible vis-à-vis the Bank; the Bank shall recognise only one owner for each share.
Instructions for Use	To manage this zones from ViPR they will have to inventory delete the volumes related to this initiators and re-ingest.
News Announcements, Reports and Research	New Zealand is particularly concerned that Timor-Leste establish a functioning Court of Appeal as soon as possible.
Patents	Then (2,4,6-trimethyl-phenyl)-acetyl chloride (0.52 g, 2.64 mmol) dissolved in dichloromethane (3 ml) was added dropwise.
Policies, Process and Procedures	By providing humanitarian assistance, UNIFIL is contributing to the improvement of the quality of life of ordinary Lebanese on a daily basis.
Sales and Marketing Material	The CX700 ensures that your enterprise storage resources are secure from both unforeseen disasters and planned events such as daily backups and application testing.
Software Strings and Documentation	AES is 128-bit Advanced Encryption Standard encryption and should be used for any network communications where security is a concern.
Support Content	Ensure cubicle jacks are properly labeled back to the wiring closet patch panel.

Datasets - Sentence Length

Too short (< 8 words) and too long (> 27 words) sentences were excluded from the dataset.

The exception is Japanese, where source texts have relatively more short segments.



2 Evaluation Methodology

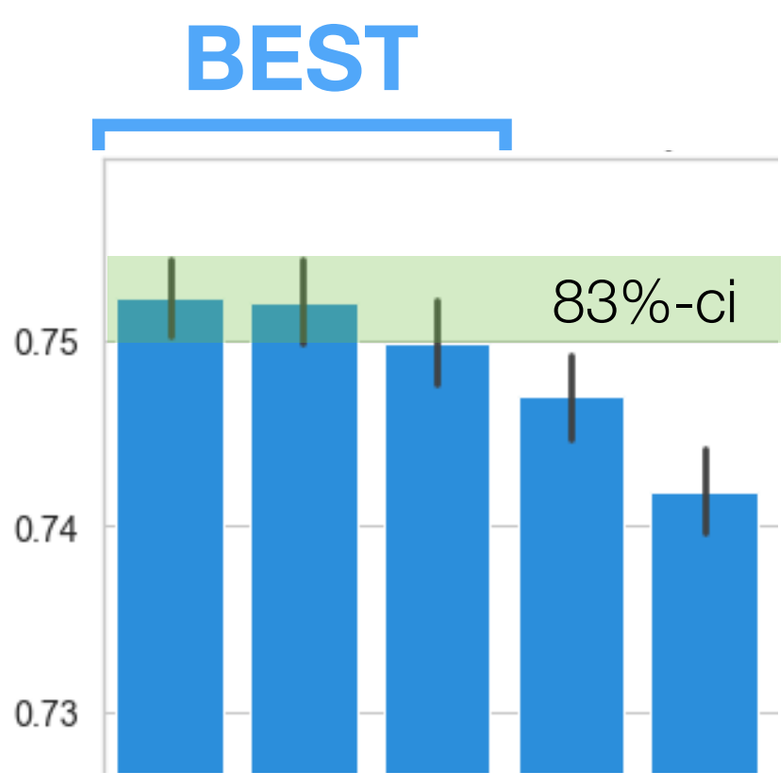
2.1 Evaluation Approach

2.2 Scores we use

2.1 Evaluation Approach

1. Rank MT engines based on a score that shows distance from a reference human translation.

2. Identify a group of top-runners (**BEST**) within a confidence interval of the leader.



Using segment-level scores averaged across the corpus and 83% confidence interval[1,2]

[1]Harvey Goldstein; Michael J. R. Healy. The Graphical Presentation of a Collection of Means, Journal of the Royal Statistical Society, Vol. 158, No. 1. (1995), p. 175-177.
[2]Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance?. *J Insect Sci*. 2003;3:34. doi:10.1093/jis/3.1.34

2.2 What scores to use?

SYNTACTIC SIMILARITY

hLEPOR (paper+code) - compares similarity of token-based ngrams. Penalizes both omissions and additions. Penalizes paraphrases / synonyms. Penalizes translations of different length.

—

SEMANTIC SIMILARITY

BERTscore (paper+code) - Analyzes cosine distances between BERT representations of machine translation and human reference **(semantic similarity)**. Does not penalize paraphrases / synonyms. May not detect factual errors (gender etc). May unreliable for terminology and synonyms in domains and languages underrepresented in BERT model.

SEMANTIC SIMILARITY

BLEURT (paper+code) - Analyzes similarity between BERT representations of machine translation and human reference **(semantic similarity)**. Does not penalize paraphrases / synonyms. Detects factual errors. May unreliable for terminology and synonyms in domains and languages underrepresented in BERT model. **Available for * => EN only.**

—

STRUCTURAL SIMILARITY

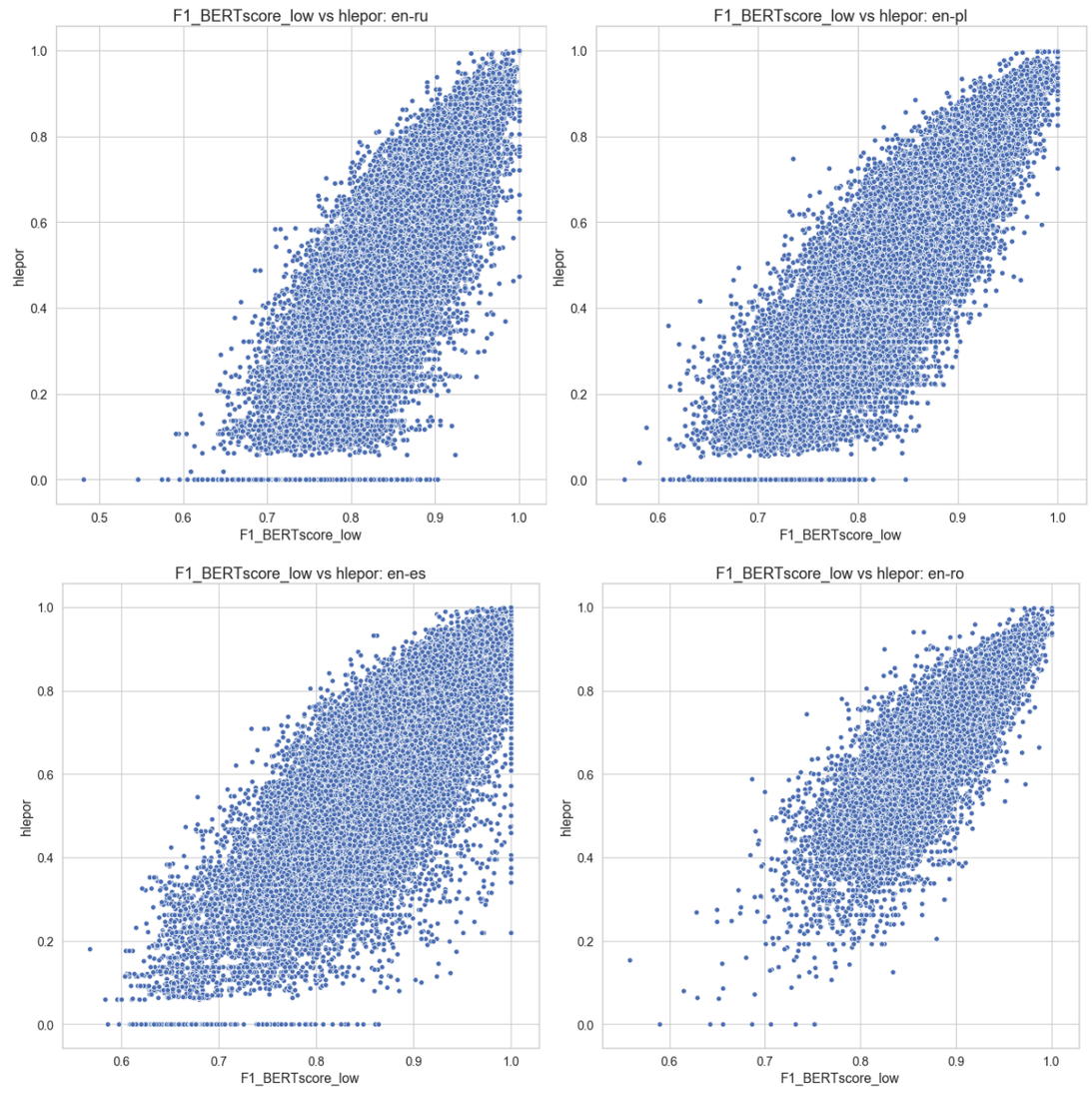
YiSi (paper+code) - analyzes structural and semantic similarity using multilingual embeddings. Only basic version for low-resource languages is available (YiSi-0). YiSi-1 uses word2vec embeddings and does not work for domain-specific texts. YiSi-2 uses BERT, but the parameters aren't published. **NOT AVAILABLE.**

Comparing hLEPOR and BERTscore

low hLEPOR **+ high** BERTscore:
> paraphrases / synonyms
> minor punctuation / tokenization issues

---

high hLEPOR **+ low** BERTscore:
> mostly doesn't exist
> a couple of segments in Japanese with an omission in reference translation



Going forward with **BERTscore**

Analyzes cosine distances between BERT representations of machine translation and human reference.

As source text and human translations are often have different case, we lowercase everything before applying BERTscore.

For Chinese, Korean and Japanese, we applied per-hieroglyph tokenization along the lines of this code.

We normalized it for every language pair.

Does not reflect absolute quality level. Not comparable across language pairs.

3 Evaluation Results

3.1 Best MT Engines per Language Pair

3.2 Best MT Engines per Industry Sector

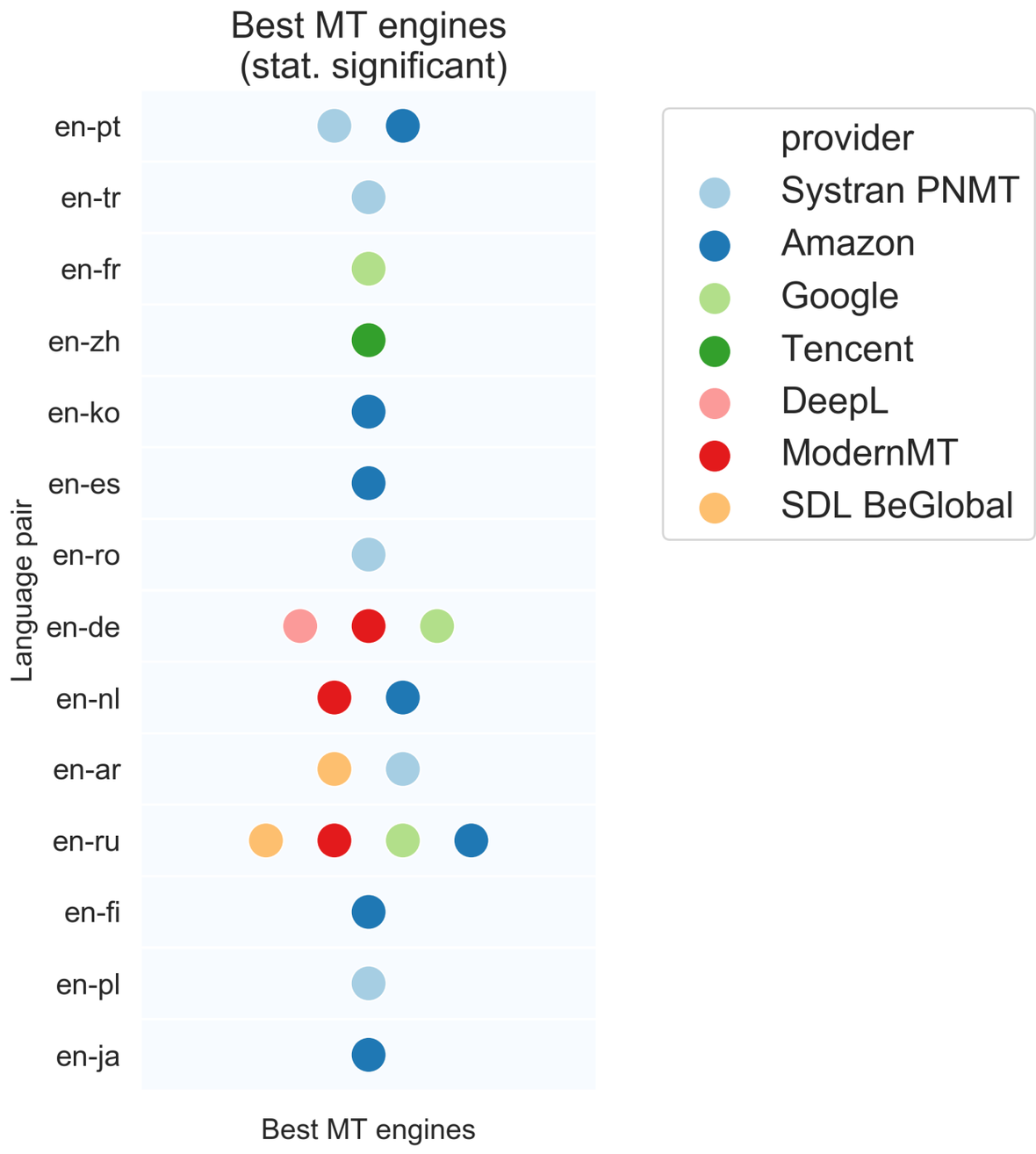
3.3 Best MT Engines per Content Type

#.4 Top-Performing MT Engines

3.1 BEST MT ENGINES PER LANGUAGE PAIR

Note the domain and content type mix is different for every language pair and is likely to influence this leaderboard a lot.

Absolute values are not shown to avoid confusion, as the score is not comparable across language pairs.



3.2 Best MT per Industry Sector

Background color shows relative quality within a language pair

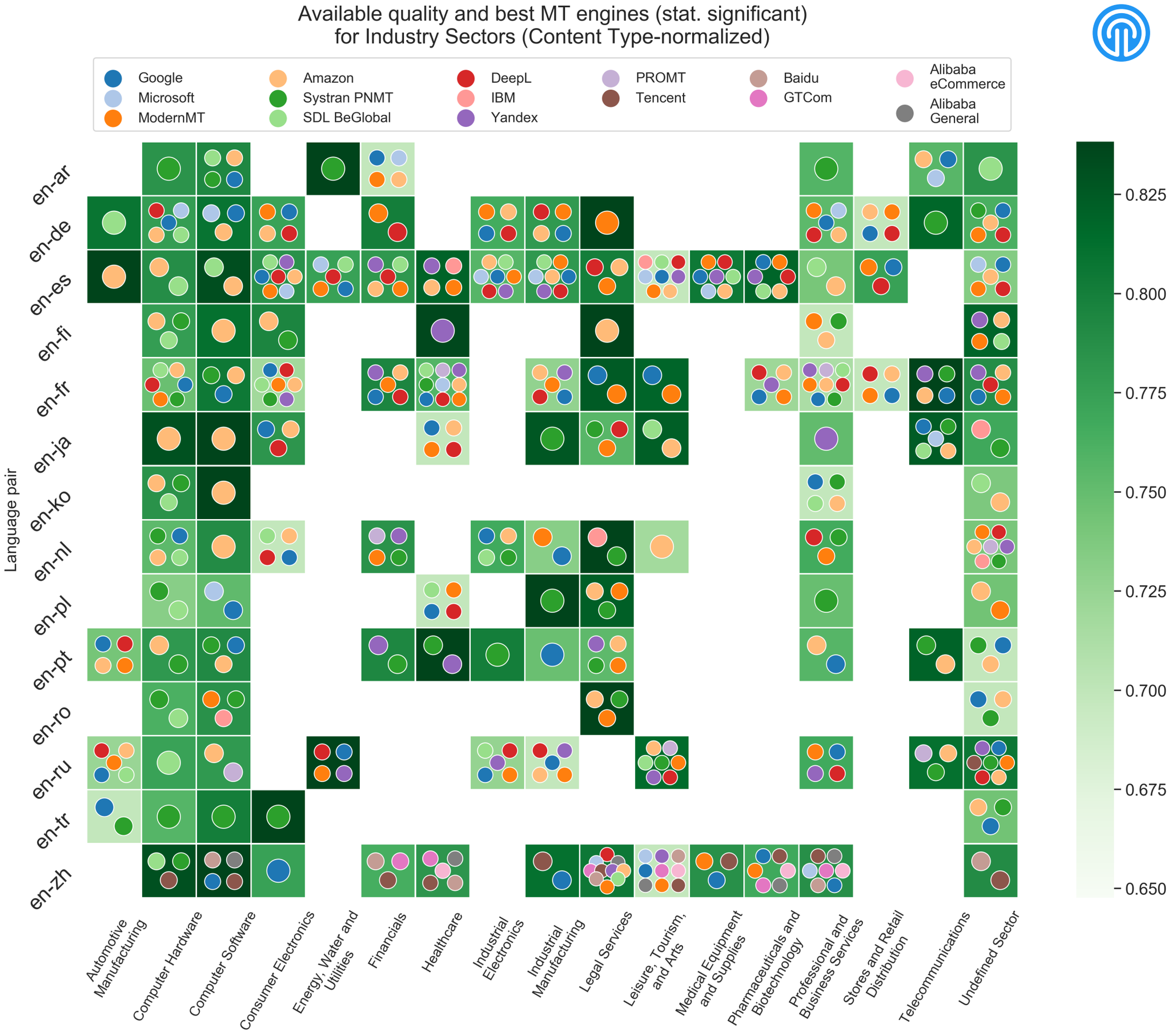
—

Standardized and scale adjusted for content type to account for different share of content types[1]

—

Not comparable across language pairs!

[1]See here for the description of standardisation and scale adjustment.



Minimal MT Coverage per Industry Sector

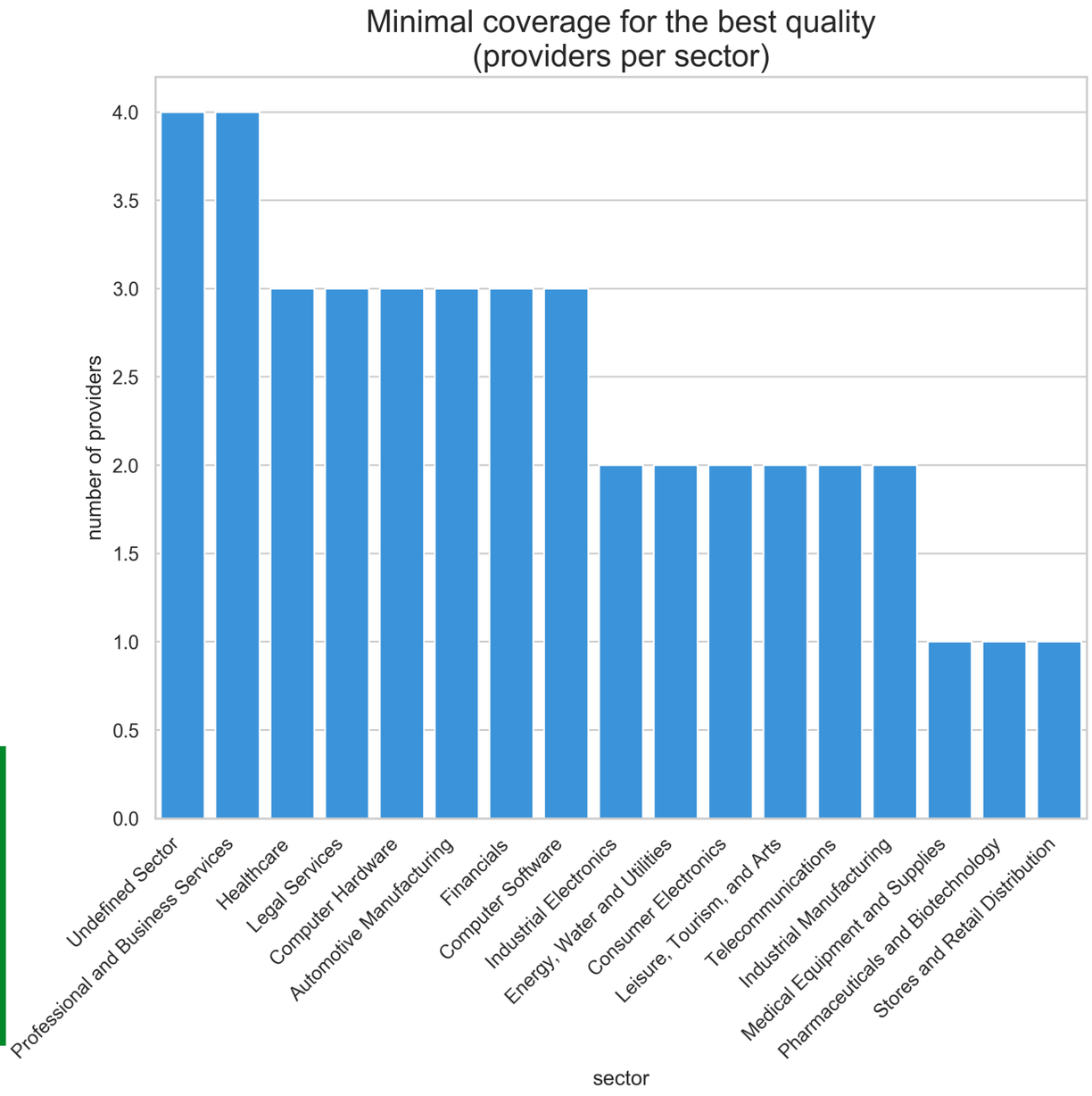
to get the best quality for all language pairs

Each of **15** different MT Providers got into the best engines in at least one Industry Sector for some language pair

7 MT providers are required to get the best quality for all language pairs and industry sectors.

However, it's enough to have **1-4 different MT providers** to get the best quality for all language pairs in any single Industry Sector (but not all of them together)

=> Any enterprise working with these language pairs should be good with 1-4 MT engines, while LSPs would need to work with at least 7



MT Complexity per Industry Sector

Computer Software, **Legal Services**, and **Telecommunications** are on higher end of the quality spectrum for every language pair

Professional and Business Services seems to be the hardest nut for MT to crack.

Many MT engines providing comparable level of quality for **Spanish, French, Chinese,** and **Russian**

Turkish, Japanese, and **Finnish** require a more careful choice of an MT engine.

Call for LSP Volunteers!

We want to go further and do test Post-Editing and LQA for each of **128** (language pair, industry sector) combinations, to estimate MT effort saving and ROI.

—

For that, we need at least two different reviewers per combination, about 10,000 words per reviewer.

—

In return, we will promote you as **Intento MT Evaluation Partner** and grant a limited **free access to the winning MT engines** via our plugins for memoQ and SDL Trados

—

Please, reach us at hello@inten.to

3.3 Best MT per Content Type

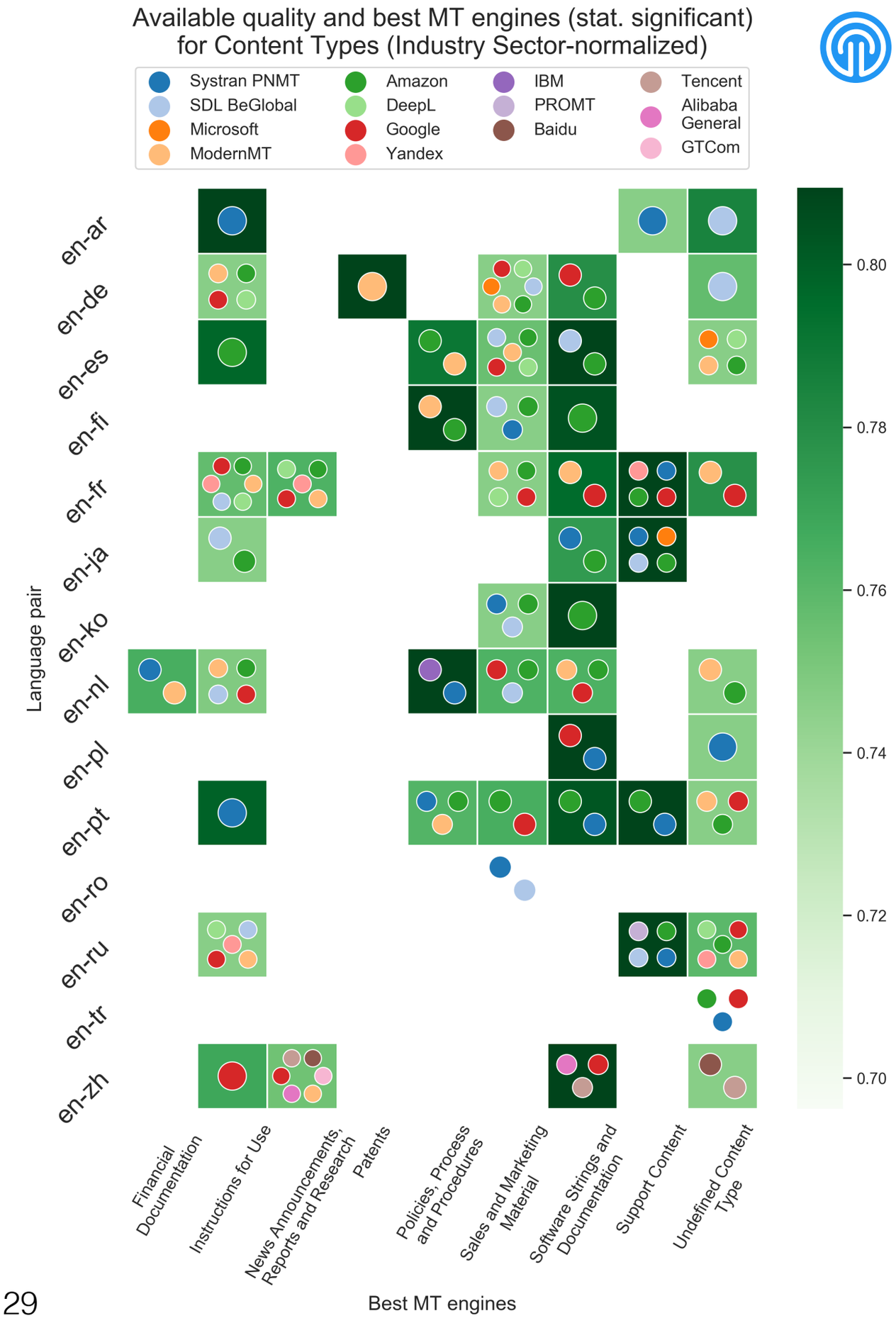
Only Content Types with >= 500 segments shown

Background color shows relative quality within a language pair

Standardized and scale adjusted for industry sector to account for different share of industry sectors[1]

Not comparable across language pairs!

[1]Similar to standardisation and scale adjustment for content types.



Minimal MT Coverage per Content Type

to get the best quality for all language pairs

Each of **14** different MT Providers got into the best engines in at least one Content Type for some language pair

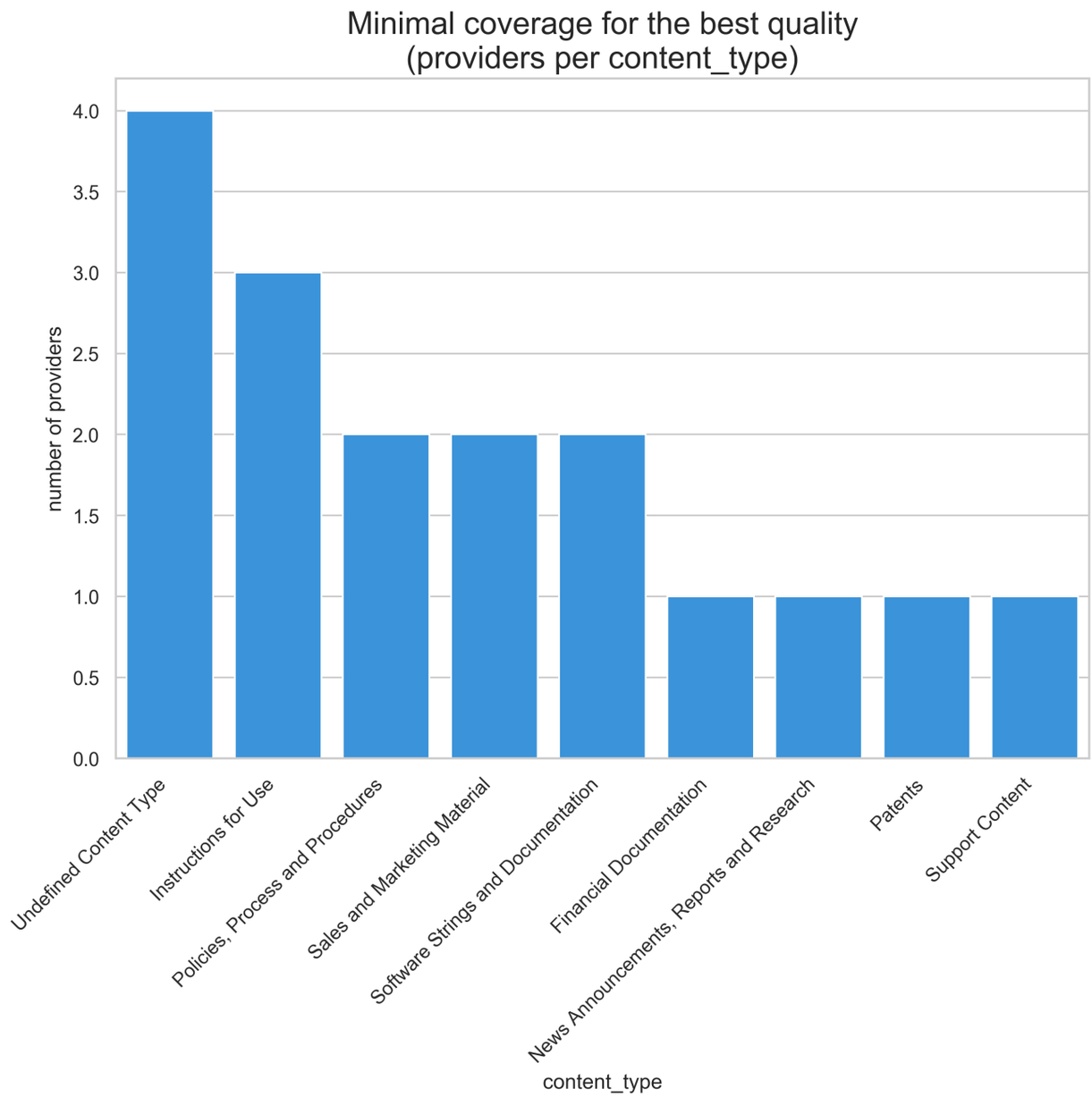
—

6 MT providers are required to get the best quality for all language pairs and content types.

—

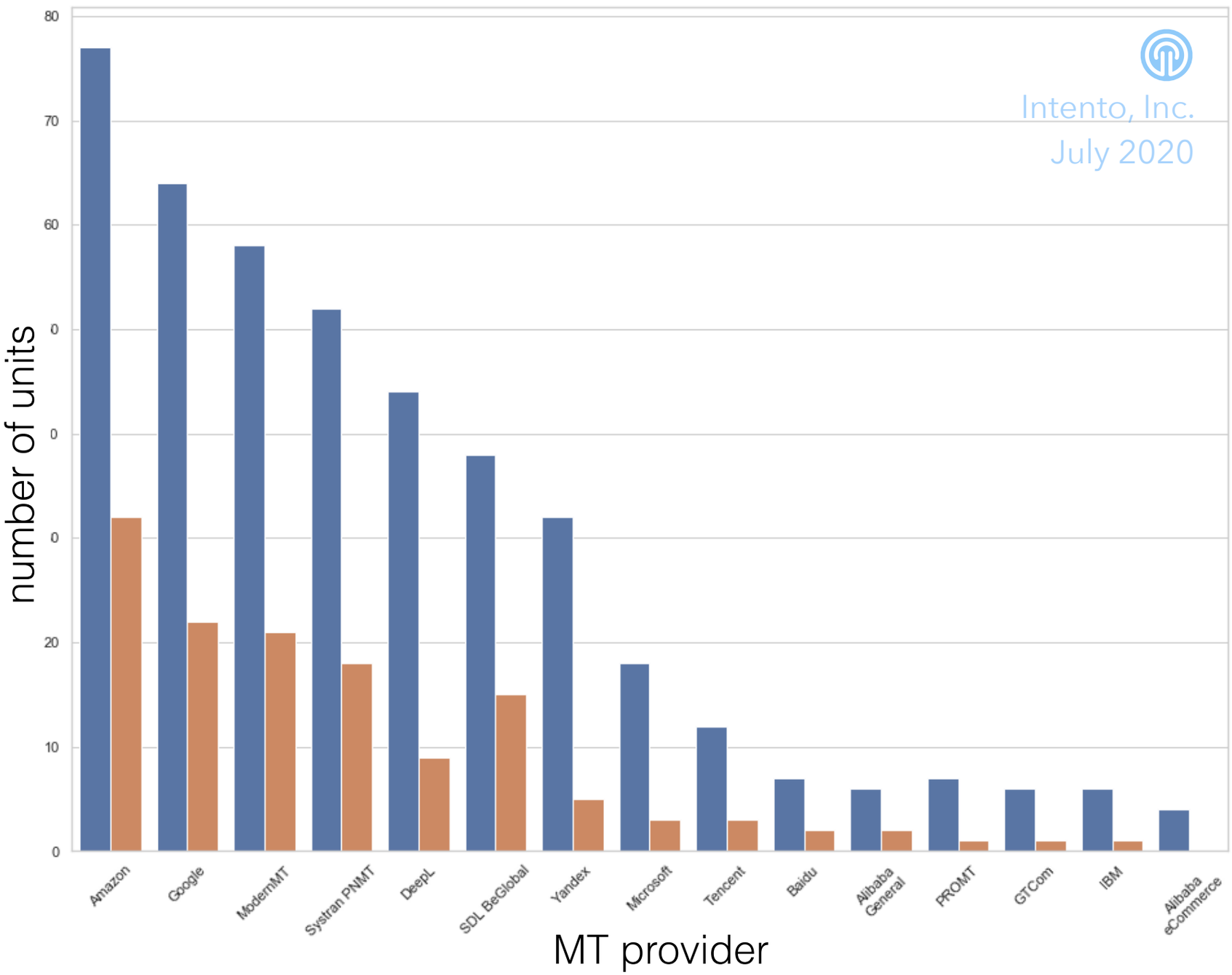
However, it's enough to have **1-4 different MT providers** to get the best quality for all language pairs in any single Content Type (but not all of them together)

=> An enterprise department working with these language pairs should be good with 1-4 MT engines, while LSPs would need to work with at least 6



3.4 TOP Performing MT Providers

across 14 language pairs, 16 industry sectors, 8 content types



pair, industry sector

Number of cases the provider got into “best”

pair, content type

Number of cases the provider got into “best”

4 Miscellaneous

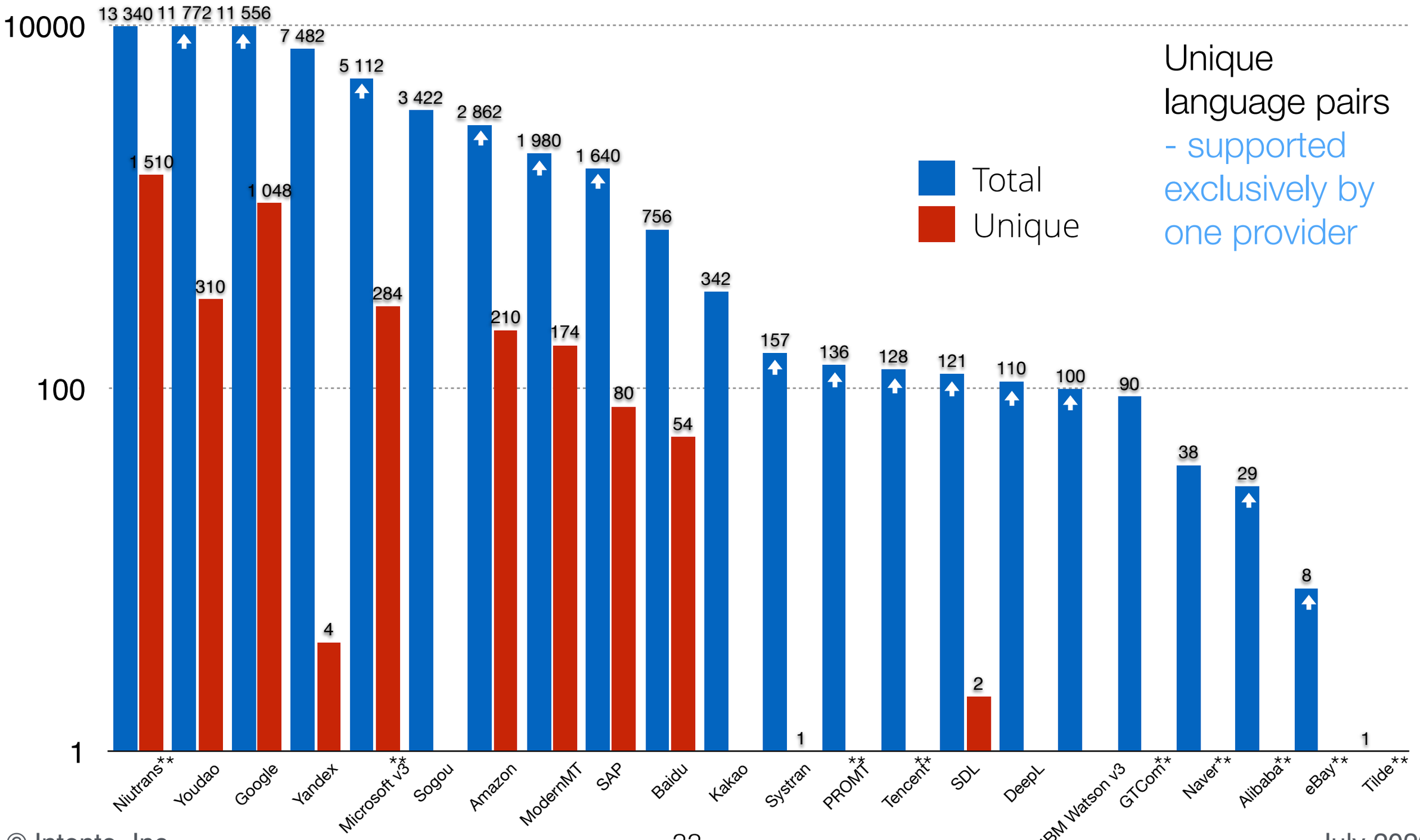
4.1 Language Support

4.2 Public Pricing

4.3 Independent Cloud MT Vendors with Stock Models

16,068 Language Pairs across all MT engines*

* where possible, we have checked via API if all language pairs advertised by the documentation are supported and removed the pairs we were unable to locate in the API.
** as advertised (not validated via API)



Unique language pairs
- supported exclusively by one provider

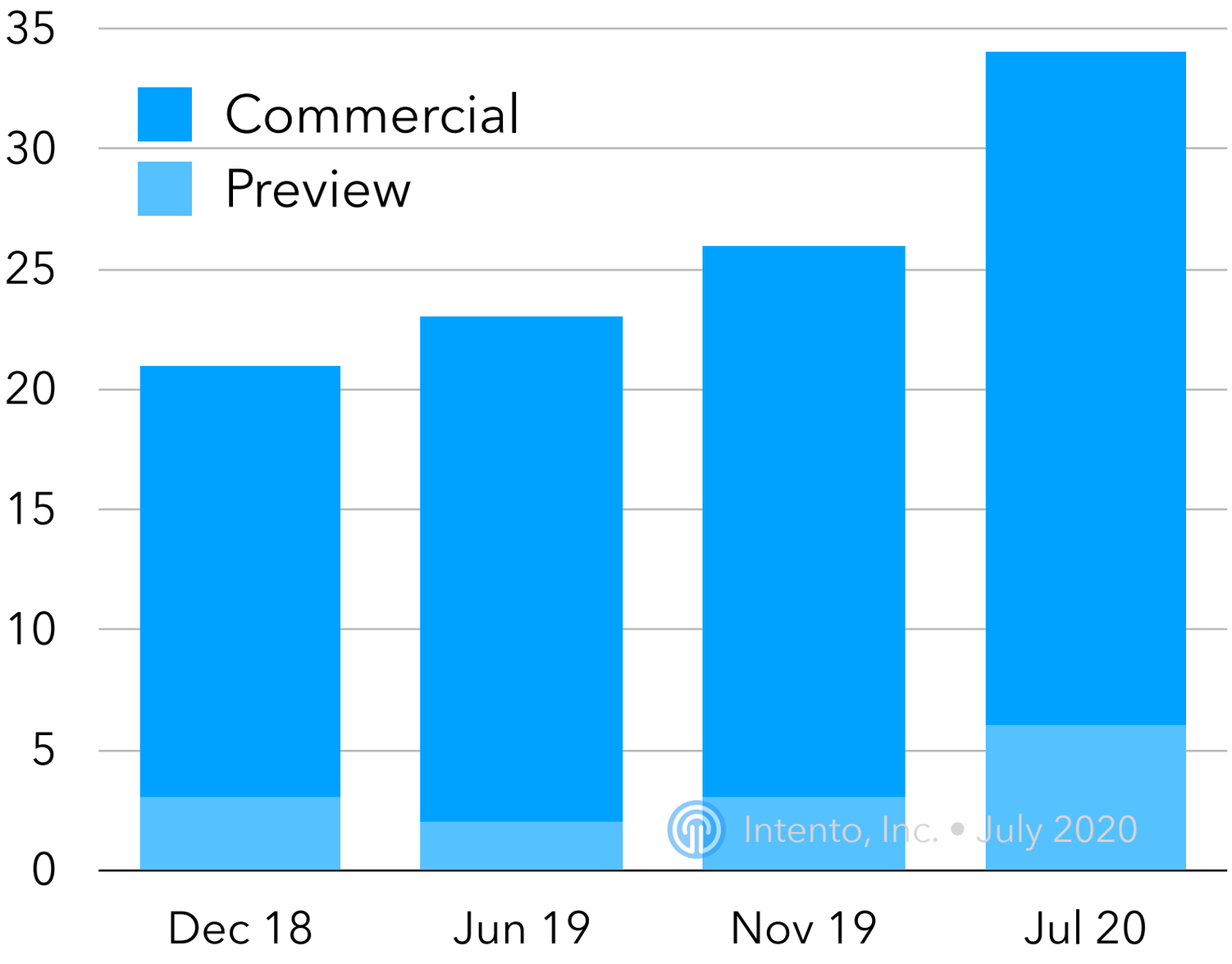
4.2 Public pricing

USD per 1M symbols***

characters per month *	Tilde	SDL BeGlobal	Cloud Translation	SAP Translation Hub	Alibaba Cloud	IBM Watson NMT/SMT	DeepL	Google Translate	SDL Language Cloud	Naver Cloud Papago NMT	ModernMT Realtime	Amazon	Systran PNMT	Microsoft NMT	GTCOM YeeCloud	Tencent	Niutrans	Baidu	Youdao	Yandex Cloud	Sogou	PROMT**	Kakao	eBay
	USD per 1,000 words			USD per 1,000,000 symbols																			free / beta	
0	no public pricing	no public pricing	no public pricing	450	33	21.4	⬇20	20	20	17.9	⬇15	15	⬇10.5	10	10	8.12	7.00	6.86	6.72	5.64	5.6 - 8.4	5		
1M					15																			
8M																						4.66		
10M																								
30M																	5.60							
32M																								
50M																								
64M																						4.53		
100M																	4.20							
128M																7.00						3.98		
200M																								
250M														8.22										
500M																								
1B														6								2.99		
1.5B																								
10B														4.5										
and more																								

* volume estimation based on 4.79 symbols per word
** +20% for some language pairs
*** freemium volumes are not shown

4.3 **More** Independent Cloud MT Vendors with pre-trained models



Commercial

Alibaba, Amazon, Baidu, CloudTranslation, DeepL, Globalese, Google, GTCom, iFlyTec, IBM, Microsoft, Mirai, ModernMT, Naver, NICT, Niutrans, Omniscien, Prompsit, PROMT, Rozetta, SAP, SDL, Sogou, Systran, Tilde, Tencent, Yandex, Youdao

Preview / Limited

eBay, Kakao, QCRI, AISA, Reverie, Tarjama

5 Conclusions

The scores are dead, long live the scores! New semantic similarity scores (e.g. BERTscore) solve the main issue of syntactic similarity score (e.g. BLEU) - dealing with alternative translations and synonyms.

Each of 15 MT engines is best at something. 7 of them are enough to get the best quality for all 15 industries and 14 language pairs we have tested. For any given industry, 1-4 engines are enough for these languages.

The highest MT quality is available for Computer Software, Legal Services, and Telecommunications, with Software Strings and Documentation, Support Content, Policies, Processes and Procedures being the easiest content types.

Everything in Professional and Business Services, as well as Instructions for Use and Sales & Marketing Content in other industries are the hardest nuts for MT to crack.

Incredible spike in language coverage: +2,000 language pairs since June 2020, many low-resource languages added (see slide 33).

MT Landscape continues to evolve: 11 more vendors offer pre-trained MT engines since the June 2019.

Intento Enterprise MT Hub

Central multi-purpose MT deployment

Universal API to all MT engines

Single MT dashboard

Smart Routing with retries and failovers

Delivers MT to many tools you use today

MT Lifecycle Management

Get your API key at inten.to

MAY BE DEPLOYED ON PRIVATE CLOUD

Intento Web Demo

End-to-End

Get a portfolio of Machine Translation engines optimal for your language pairs, domains, and available training data.

Fast and Safe

4-5 weeks from assorted TMs and glossaries to winning MT engines with ROI estimation for Post-Editing and Real-Time Machine Translation.

Trusted

We run 15-20 MT Procurement projects per month for global retail, travel, and technology companies under strict Security, Quality and Data Protection requirements. ISO 27001 certified.

REACH US

at hello@inten.to

Intento Plugins and Connectors

Microsoft Office (Outlook, Word, Excel)

—

Google Chrome and Microsoft Edge (extension)

—

memoQ (included in 9.4, also private plugin)

—

SDL Trados (SDL AppStore)

—

XTM (XLIFF API Connector)

—

MateCat (private plugin)

—

Any Enterprise TMS via XLIFF connector.

—

Miss some connector? Reach us at hello@inten.to!

STATE OF THE MACHINE TRANSLATION

STOCK* MT MODELS

by Intento (https://inten.to)

July 2020

Intento, Inc.
hello@inten.to
2150 Shattuck Ave
Berkeley CA 94704

* commercially available pre-trained MT models

data provided by TAUS the language data network

Appendix A

Overall performance of the MT services for an Industry Sector across many Content Types is computed in the following way:

1. [Standardisation] We compute Content-Type-standardized score (or z-score) for each Industry Sector.
2. [Scale adjustment] We restore the original scale by multiplying z-score by the score standard deviation for this sector and adding the score mean for this sector.

$$score^{STD} = \mu_{pair,sector} + \sigma_{pair,sector} * \frac{score - \mu_{pair,sector,content}}{\sigma_{pair,sector,content}}$$

scale adjustment *z-score*