Intento

Independent Multi-Domain Evaluation of Machine Translation Engines

Part 1: Automatic semantic similarity scoring

In partnership with

TAUS
the language data network

https://inten.to

Disclaimer

The MT systems used in this report were accessed from June 4 to August 16 2021. Some of these systems may have changed since this time.

The data originates from several large companies, and is available for purchase via TAUS Data Marketplace. MT providers could have had access to such data in the past to use for training their models.

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.

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

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

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

All third-party trademarks, registered trademarks, product names, and company names or logos mentioned in the Report are the property of their respective owners, and the use of such Third-Party Trademarks inures to the benefit of each owner. The use of such Third-Party Trademarks is intended to describe the third-party goods or services and does not constitute an affiliation by Intento and its licensors with such company or an endorsement or approval by such company of Intento or its licensors or their respective products or services.

Executive **Summary**

The **MT market is accelerating. 13 more vendors** offer pre-trained MT models since August 2020, plus there are several **open-source** pre-trained MT engines available. We have evaluated **29 MT engines - 14 more than a year ago!**

Unprecedented language coverage: **99,760 language pairs** across all MT engines. It was just 16K a year ago! The main contributors are **Niutrans** with their 88K language pairs and **Alibaba** with 20K.

19 MT engines are among the statistically significant leaders for **7** industry sectors and **13** language pairs. **9** MT engines provide minimal coverage for all language pairs and industries, **1-4** per industry sector.

Many engines perform best with English to **Spanish**, **Russian**, and **Chinese**. **Legal**, **Financial**, and **Healthcare** require a careful choice of MT vendor, as relatively few perform at the top level. Despite having several comparable MT engines per language pair, **Education** shows relatively low scores, which may indicate the importance of customization in this domain.

Open-source engines perform in the 2nd tier of commercial systems, except for **en-es** (on par with top-tier systems) and **en-ko** & **en-ja** (much worse than commercial systems).

New scores on the block! This time, we have added COMET and PRISM, and checked how they perform compared to BERTScore.

About

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

As we demonstrate in this report, the Machine Translation landscape is both complex and dynamic. Models from nine different vendors are required to get the best quality across popular language pairs and there is a 90x difference in price.

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 MT Hub for Enterprise

Intento MT Hub and MT Studio for Enterprise

Make the most of the Machine Translation landscape with advanced tools from Intento.

MT Hub

Integrate AI/ML models from many vendors into your business processes, choosing the best-fit combination for every use case.

- Localization
- Office Productivity
- Software Development
- Customer Service
- Community & Marketing

MT Studio

Train, evaluate, and improve the best-fit machine translation model from a single interface.

- Data Cleaner
- Trainer
- MT Customization Analysis
- Scoring
- LQA tools
- Analysis tools
- Routing Designer
- Glossary Management
- Feedback Management

Learn more at https://inten.to

Book a live demo

Intento

Overview

1. Datasets

2. Evaluation methodology

3. Evaluation results

4. Miscellaneous

5. Key conclusions

29 Machine Translation Engines

13 Language Pairs

7 Industry sectors

Machine Translation Landscape

Generic stock models

AISA, Alibaba, Amazon, AppTek, Baidu, DeepL, eBay, Elia, Fujitsu, Globalese, Google, GTCom, IBM, iFlyTec, Kakao, Kingsoft, Lesan, Lindat, LingvaNex, Microsoft, Mirai, ModernMT, Niutrans, Naver, Kawamura (NICT), NTT, Omniscien, Pangeanic, Process9, Prompsit, PROMT, Rozetta, RWS, SAP, Sogou, Systran, Tencent, Tilde, Vicomtech, XL8, Yandex, YarakuZen, Youdao

Auto domain adaptation

Amazon, Globalese, Google, IBM, Kantan, Microsoft, ModernMT, Omniscien, RWS, Systran

Vertical stock models

Alibaba, Baidu, CloudTranslation, RWS, Microsoft, Omniscien, PROMT, SAP, Systran

Custom terminology support

Amazon, Baidu, Google, IBM, Microsoft, Rozetta, RWS, Systran, Yandex

Manual domain adaptation

Alibaba, AppTek, Baidu, CloudTranslation, Omniscien, Pangeanic, Prompsit, PROMT, RWS, Systran, Tilde, Yandex

Machine Translation Engines

Evaluated in this study

Customization options: none · TM · glossary · both

Engine	Product	Customization options
Alibaba Cloud	eCommerce MT	none
Alibaba Cloud	General	none
Amazon	Translate	both
Apptek	Neural Machine Translation	none
Baidu	Translate API	glossary
DeepL	API	glossary
Elia	Elhuyarren itzultzaile automatikoa	none
Globalese	MachineTranslation	TM
Google Cloud	Advanced Translation	both
GTCom	YeeCloud MT	none
IBM Watson	Language Translator	both
M2M-100-1.2B	Open-source model	none
M2M-100-418M	Open-source model	none
mBART50-EN2M	Open-source model	none
mBART50-M2M	Open-source model	none
Microsoft	Translator Text	both
ModernMT	Realtime	TM
Naver	Papago NMT Commercial	none
Kawamura	NMT powered by NICT	none
OPUS MT	Open-source model	none
Pangeanic	Machine Translation API	none
PROMT	Cloud API	none
Rozetta T-400	Machine Translation API	none
Systran	PNMT	both
Tilde	Machine Translation API	none
Tencent Cloud	TMT API	none
Yandex	Translate API	TM
Youdao	Cloud Translation API	none
XL8	Machine Translation	none

Intento

1

Datasets

Intento

Datasets — Origin

The datasets are provided by TAUS – one-stop language data shop

Every element has the following attributes:

- Source text
- Reference human translation
- Industry sector

Data samples to reproduce this study are available from TAUS and Intento

Datasets — Cleaning

The first cleaning was performed by TAUS

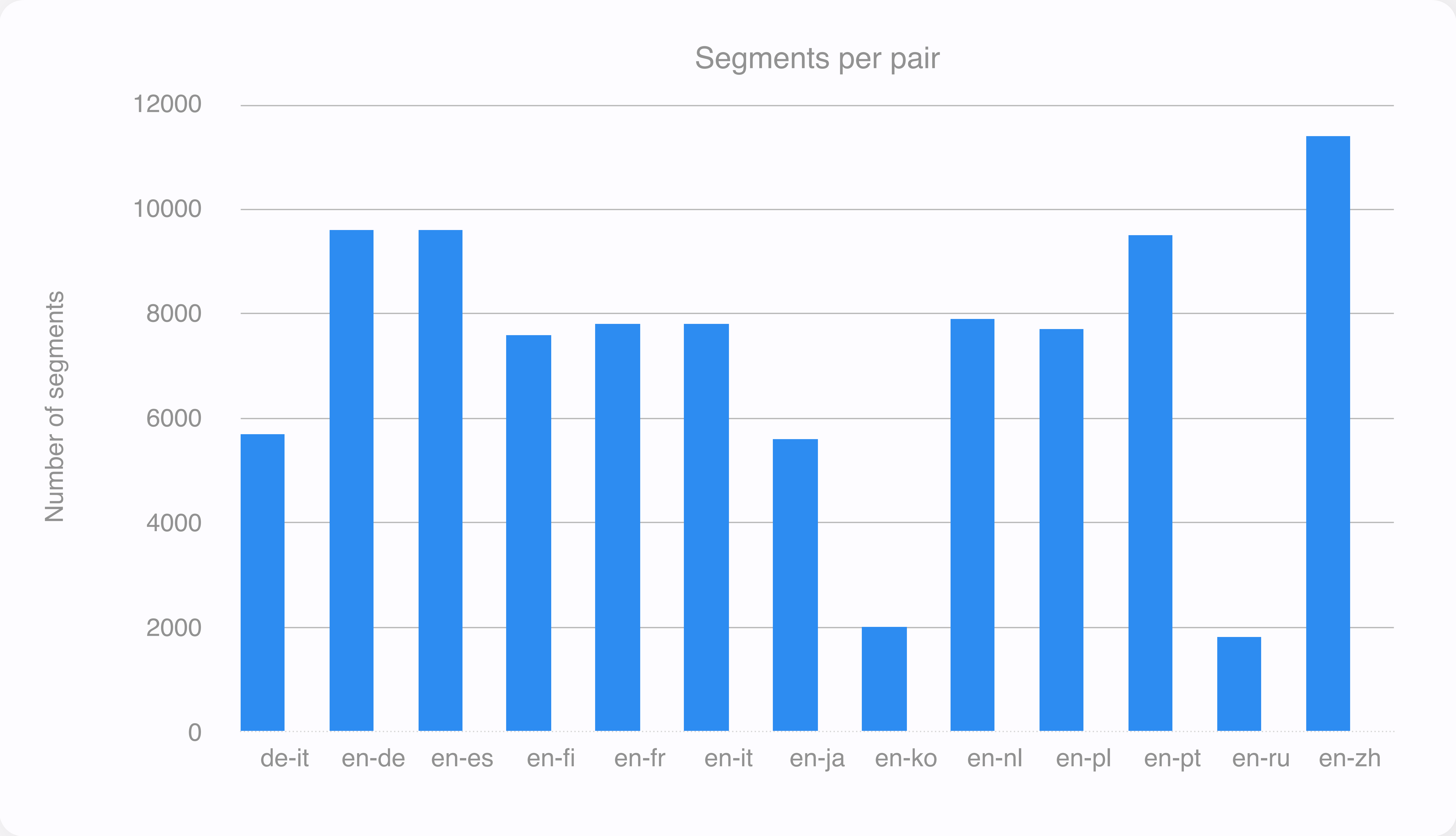
Additional cleaning by Intento:

- Removed duplicates
- Removed segments under 4 words
- Removed mistranslations using Automated Translation Quality Estimation (based on multilingual embeddings)

Intento

Datasets — Language Pairs

13 language pairs, selected based on the availability of **~2000 segments** with high-quality translations for several industry sectors.



Intento

Datasets — Industry Sectors

- **1–7** industry sectors per language pair
- **1,700–3,800** segments per language pair per industry sector
- For Russian and Korean, only Financial domain is available

Available resources, sector x pair

sector \ pair	de-it	en-de	en-es	en-fi	en-fr	en-it	en-ja	en-ko	en-nl	en-pl	en-pt	en-ru	en-zh
Education	1751	1846	1798	1715	1784	1894	3796		1947	1778	1836		1914
Entertainment													1901
Financial	1874	1918	1915	1907	1932	1828		2000	1983	1817	1897	1830	1827
General							3818						
Healthcare		1968	1972	1886	1988	1959			1981	1986	1941		1886
Hospitality		1842	1845				3314				1839		1858
Legal	1892	1958	1983	1806	1947	1973			1986	1974	1980		1948

Intento

Content Samples

Industry Sectors

Education
“She attended Lane Cove Public School before going to high school at SCEGGS Darlinghurst.”

Finance
“Credit ratings have regulatory value for regulated investors, such as credit institutions, insurance companies and other institutional investors.”

Healthcare
“The effects of Pramipexole Teva may be altered or side effects may occur if you are also taking other medicines.”

Hospitality
“Looks like he's staying at a motel in Hamakua, and he just booked a return flight back to L.A. that leaves in a few hours.”

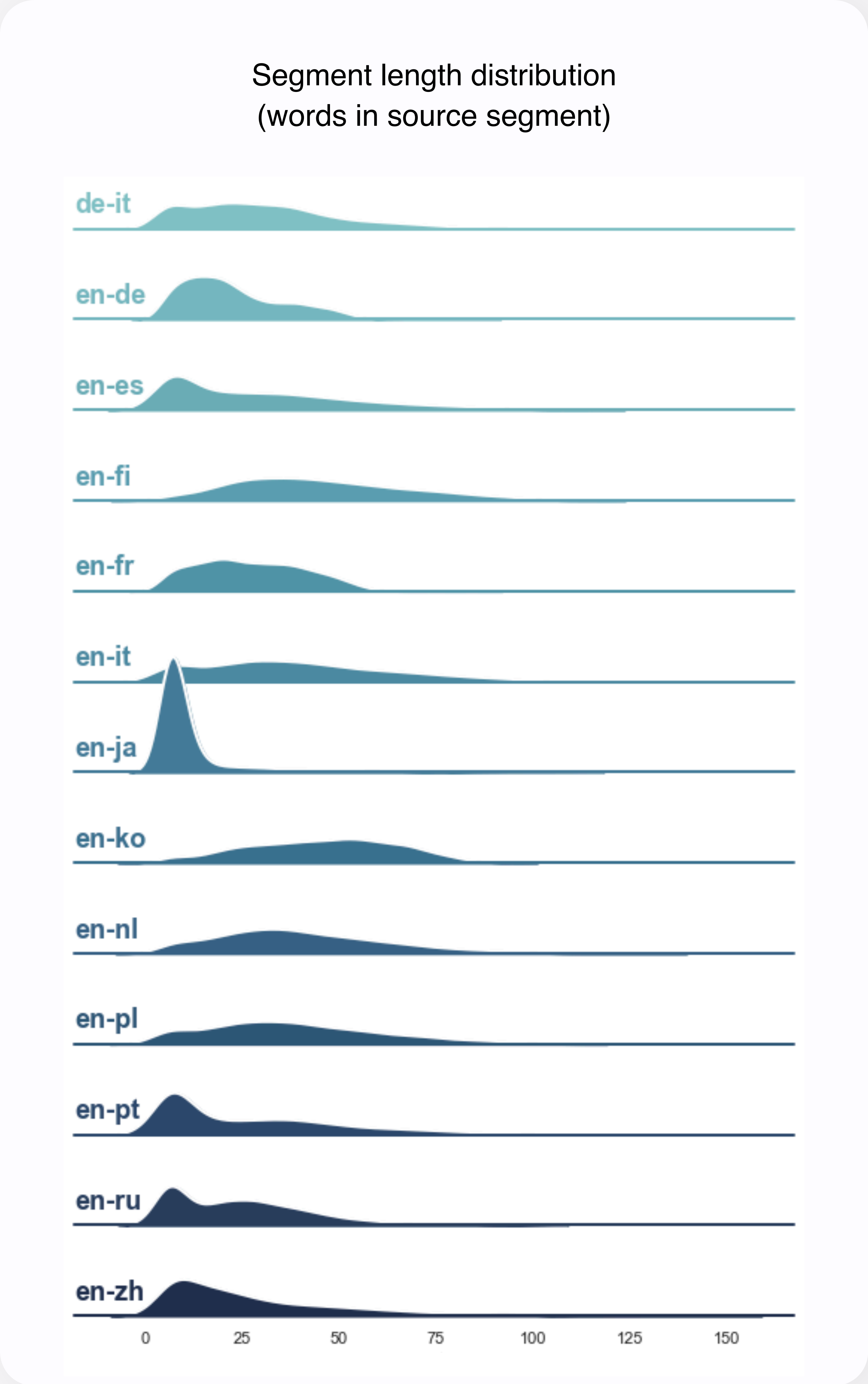
Legal
“The Court therefore considered it necessary to examine whether that exclusive right can be justified on the basis of Article 86(2).”

Entertainment
“The racks are full of magazines reporting on the lives of TV and film stars, athletes, singers, musicians, famous politicians, and foreign royalty.”

General
“The other day I stopped at a secondhand bookstore on my way home from school and happened to find a book I had been looking for for a long time.”

Datasets — Sentence Length

- Too short (< 4 words), and were excluded from the dataset.
- The exception is Japanese, where source texts have relatively more short segments.



2

Evaluation Methodology

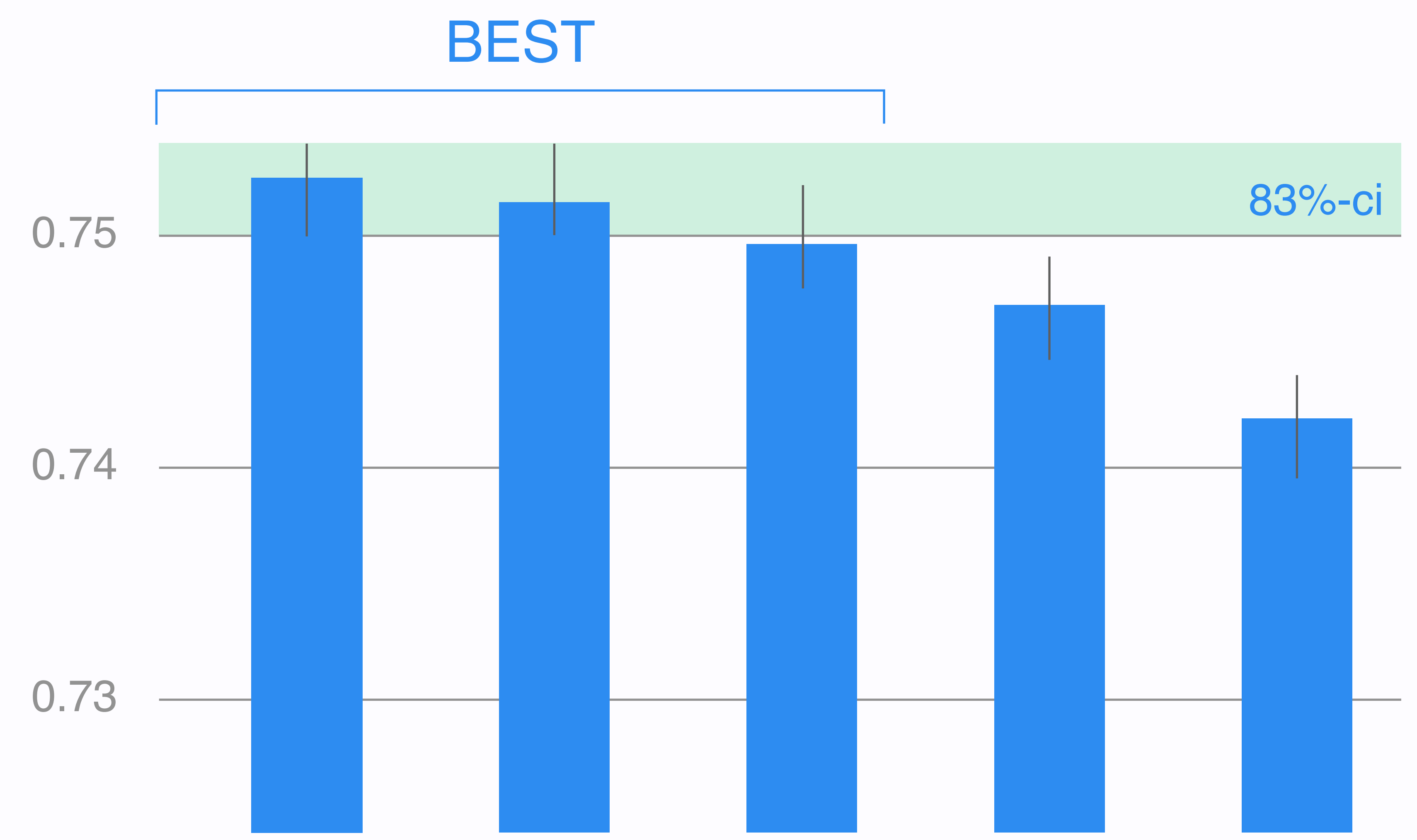
Intento

2.1 Evaluation Approach

1. Rank MT engines based on a score showing 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 an 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 be unreliable for terminology and synonyms in domains and languages underrepresented in BERT model.
SYNTACTIC SIMILARITY	**TER** paper + code	Measures the number of edits (insertions, deletions, shifts, and substitutions) required to transform a machine translation into the reference translation. Penalizes paraphrases/synonyms. Penalizes translations of different length.
SEMANTIC SIMILARITY	**PRISM** paper + code	Evaluates machine translation as a paraphrase of a human reference translation. Penalizes both fluency and adequacy errors. Does not penalize paraphrases/synonyms. N/A for Korean.
SEMANTIC SIMILARITY	**COMET** paper + code	Predicts machine translation quality using information from both the source input and the reference translation. Achieves state-of-the-art levels of correlation with human judgement. May penalize paraphrases/synonyms.

Intento

Choosing the Score

- We decided to decommission n-gram based scores (**hLEPOR**, **TER**, **BLEU**) as we observe an increasing amount of good paraphrases from MT, and they all received low scores.

- We cannot use **PRISM** for the purposes of this report as we observe unstable behavior, with translations similar to the reference getting scores lower than some of the imperfect paraphrases, making comparing the mean scores problematic for high-performing engines. Also, it does not penalize non-translations and is not available for Korean.

- A choice has to be made between **BERTScore** allowing omissive paraphrasing, and **COMET** penalizing context-depending alternative translations. We have decided to go with **BERTScore** for this report, as it may be more relevant in reflecting the understandability of the translations.

- We also provide results for **COMET**, as there's enough evidence in the literature to suggest a greater correlation with linguistic quality, which may be important for MTPE and some other use-cases.

See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix A

Going Forward with BERTScore

- Analyzes cosine distances between BERT representations of machine translation and human reference.
- Source texts and human translations often have different formatting, so we lowercase everything before applying BERTScore.
- For every language pair, we have normalized the BERTScore to fit [0,1] interval.
- Does not reflect absolute quality level. Not comparable across language pairs.
- Our version of the BERTScore is available for Intento customers.
Reach us to learn more.

See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix A
See the analysis for COMET and PRISM in Appendix B and Appendix C

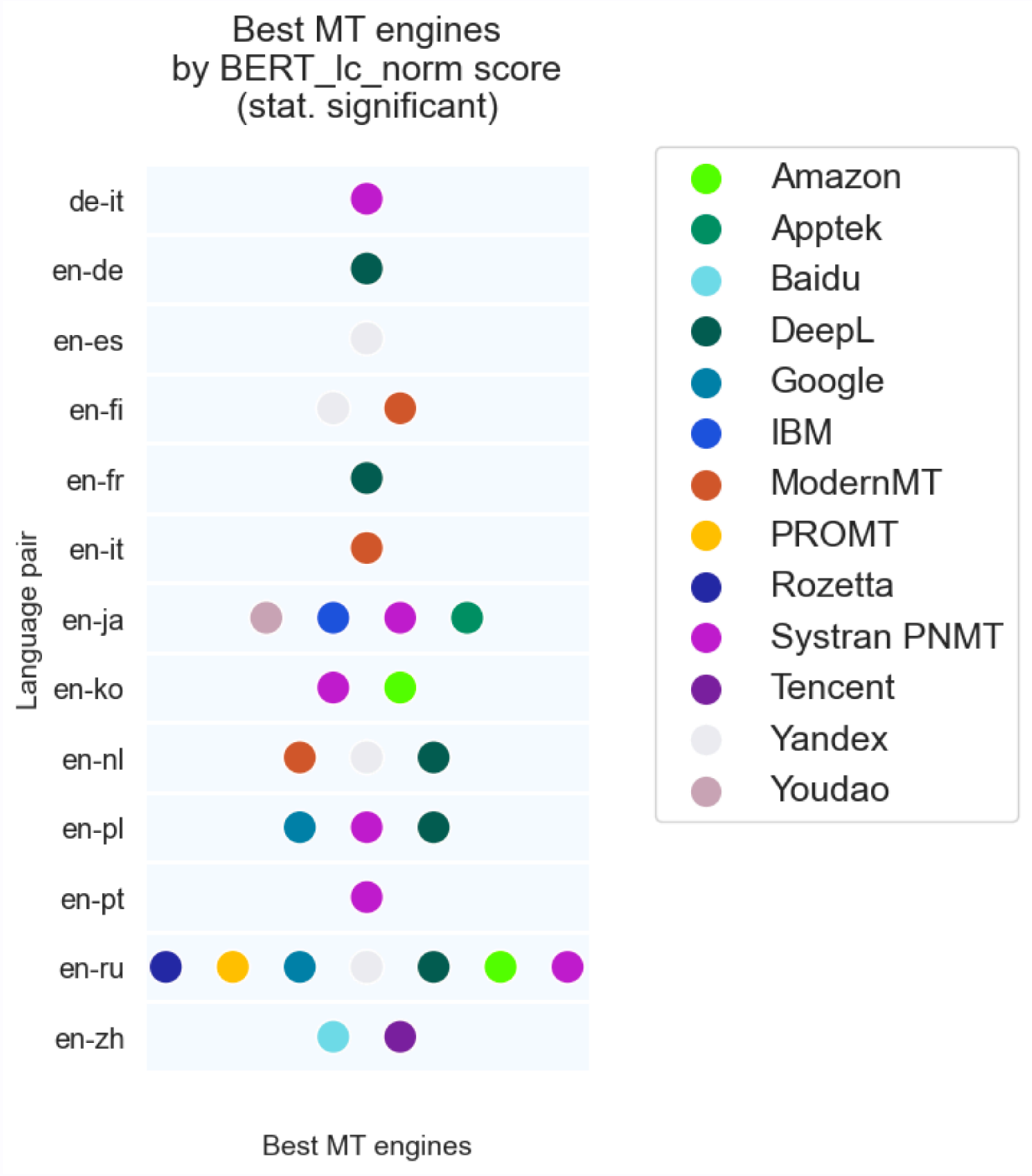
3

Evaluation Results

Intento

3.1 Best MT Engines per Language Pair (BERTScore)

- 13 MT engines are among the statistically significant leaders for 13 language pairs
- 5 MT engines cover the best scores for all 13 languages: DeepL, Systran, Yandex, ModernMT, and Baidu or Tencent
- Absolute values are not shown to avoid confusion, as the score is not comparable across language pairs.
- The domain and content type mix is different for every language pair (see the next slide) and largely influences this leaderboard.

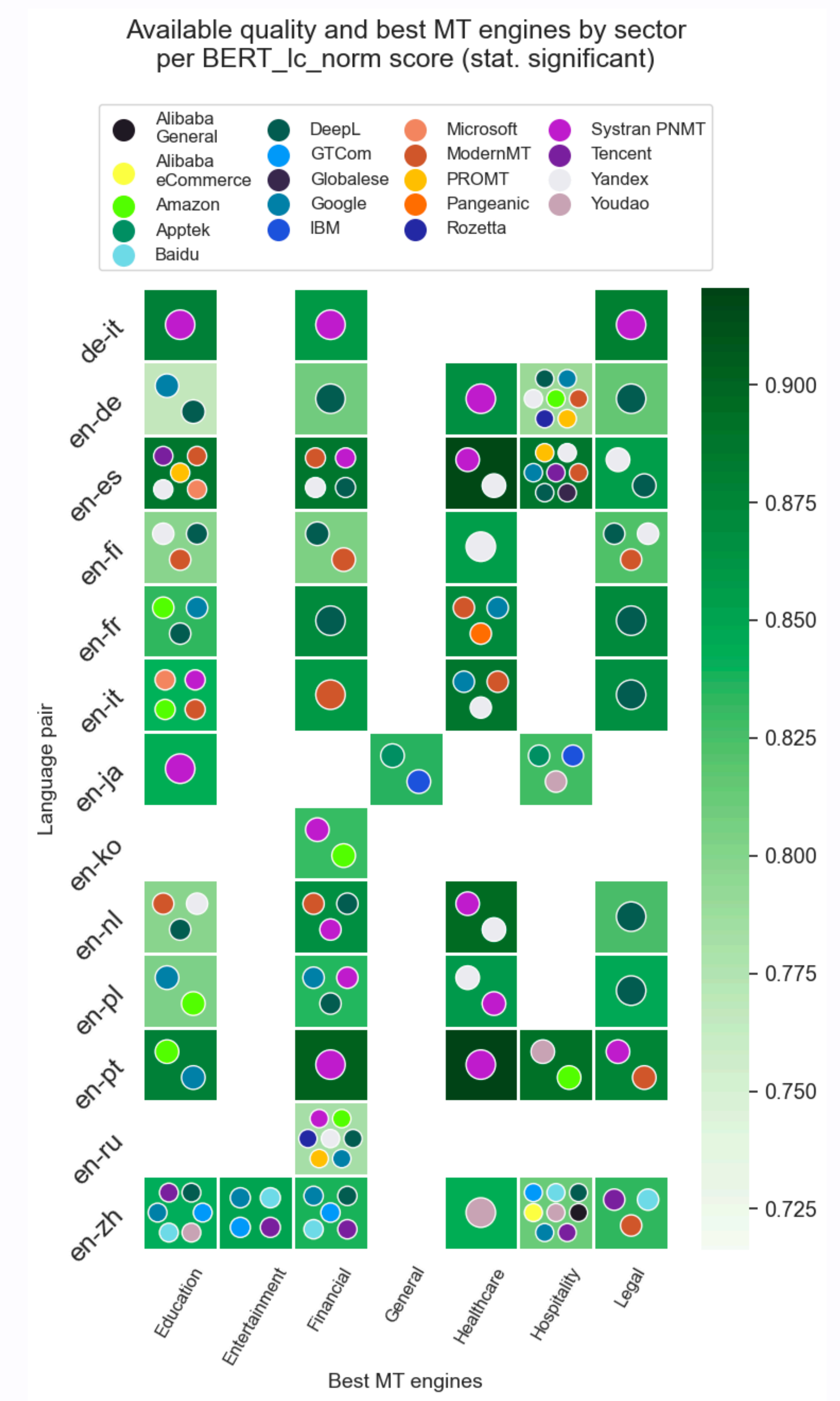


October, 2021

Intento

3.2 Best MT Engines per Industry Sector

- **19** MT engines are among the statistically significant leaders for **7** industry sectors and **13** language pairs.
- **9** MT engines provide minimal coverage for all language pairs and industries, **1-4** per industry sector.
- Many engines perform best with English to **Spanish**, **Russian**, and **Chinese**.
- **Legal**, **Financial**, and **Healthcare** require a careful choice of MT vendor, as few perform at the top level.
- Despite of having several comparable MT engines per language pair, **Education** shows relatively low scores, which may indicate the importance of customization in this domain.
- It appears that **Systran** is the only engine that translates German to Italian without the pivot through English.

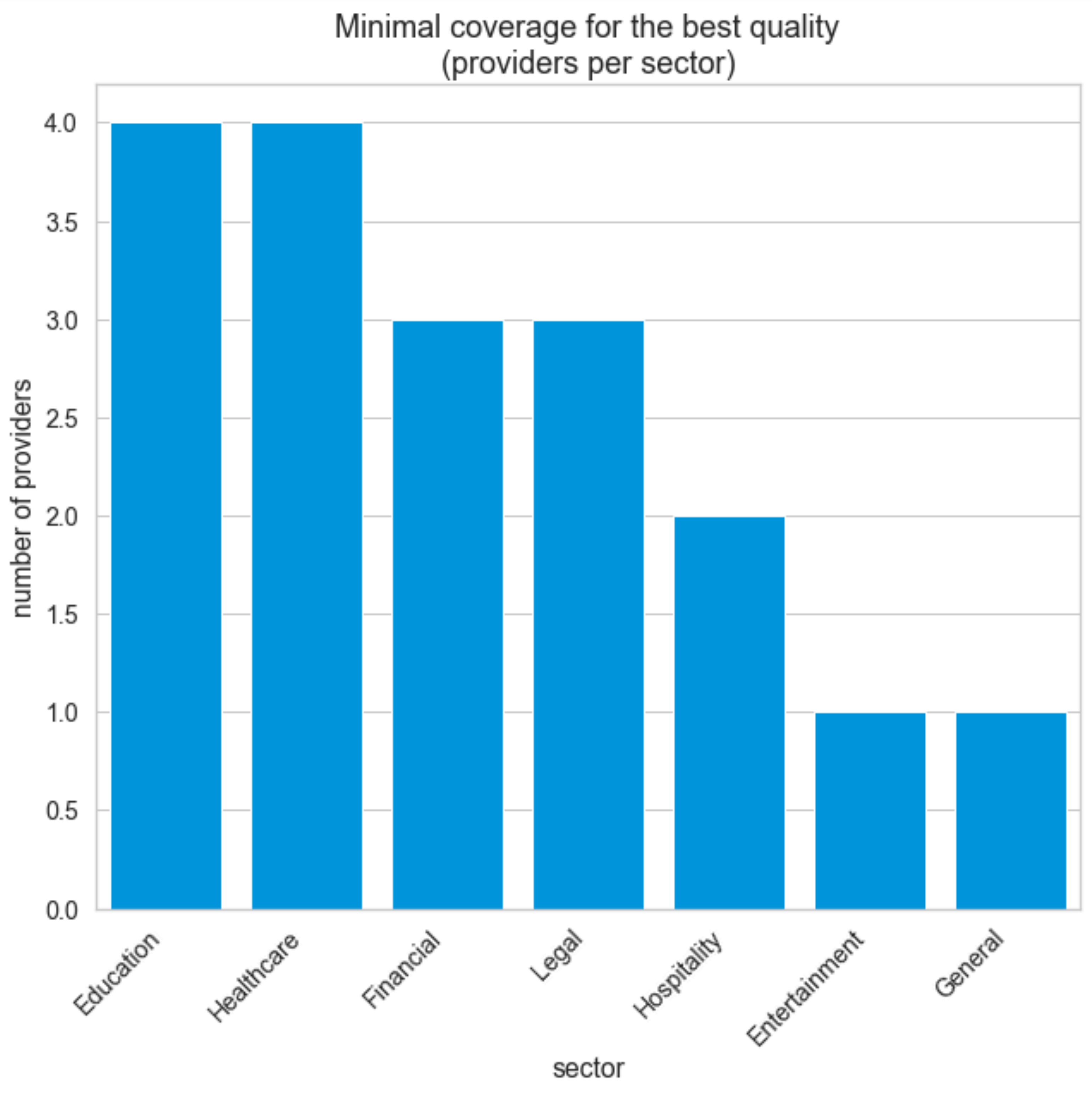


 October, 2021

Intento

3.3 Possible Minimal Coverage

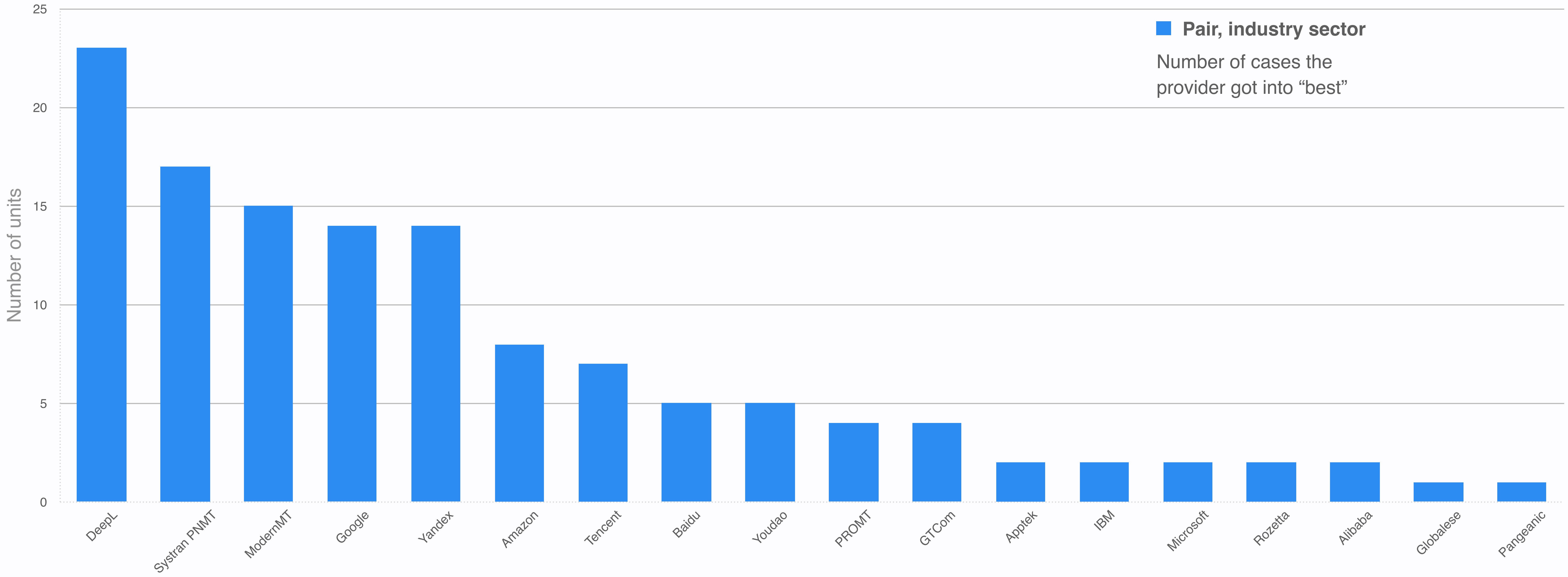
- **Education:** **Systran** (de-it, en-it, en-ja), **DeepL** (en-de, en-fi, en-fr, en-nl, en-zh), **Amazon** (en-pl, en-pt), **Microsoft** (en-es)
- **Healthcare:** **Systran** (en-de, en-es, en-nl, en-pl, en-pt), **Yandex** (en-fi, en-it), **Youdao** (en-zh), **Google** (en-fr)
- **Financial:** **Systran** (de-it, en-es, en-ko, en-nl, en-pl, en-pt, en-ru), **DeepL** (en-de, en-es, en-fi, en-fr, en-it), **ModernMT** (en-it)
- **Legal:** **Systran** (de-it, en-pt), **DeepL** (en-de, en-fi, en-fr, en-nl, en-pl), **Baidu** (en-zh)
- **Hospitality:** **Baidu** (en-zh)
- **General:** **Google** (en-ja)



For every industry sector, we provide one of the possible minimal coverages.

3.4 TOP Performing MT Providers (BERTScore)

Across 13 language pairs, 7 industry sectors



4

Miscellaneous

Intento

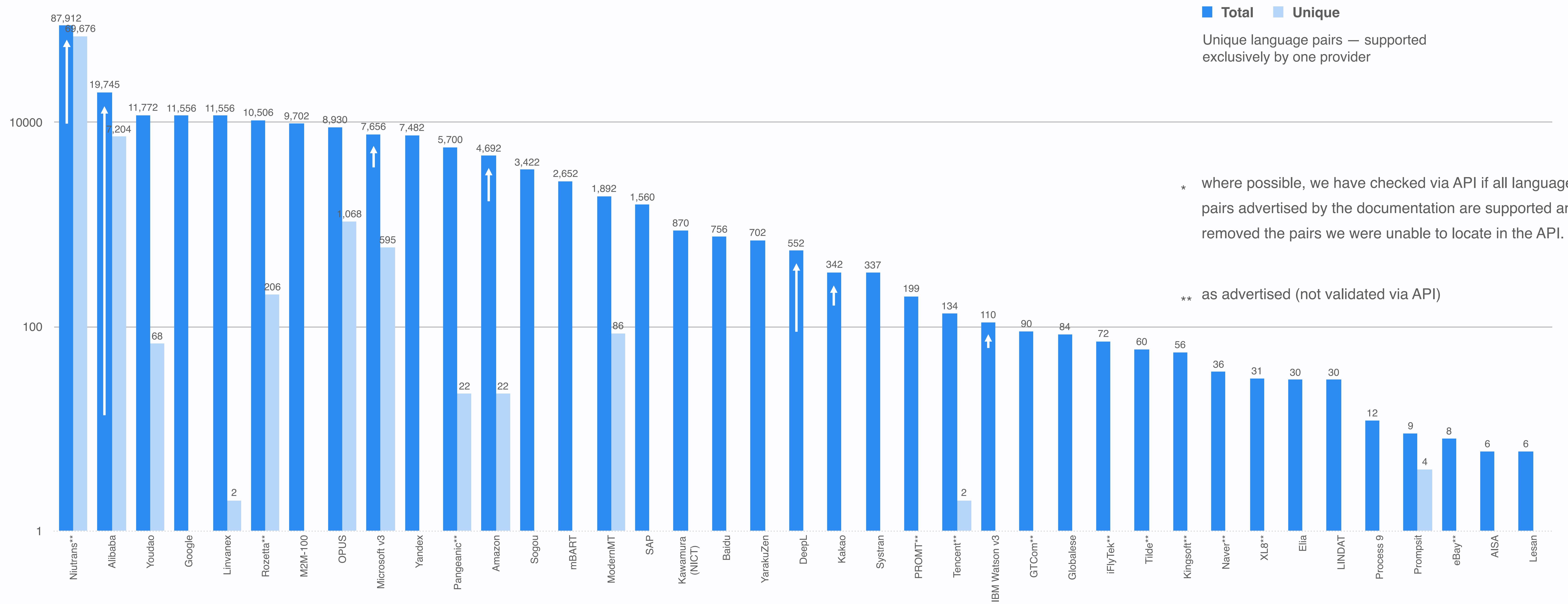
99,760 Language Pairs Across All MT engines*

From 16,068 in August’20 to 99,760 in August’21

Significant growth for Niutrans and Alibaba

Also growth for Microsoft, Amazon, DeepL, Kakao, IBM

Many new niche MT providers with few languages



* 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)

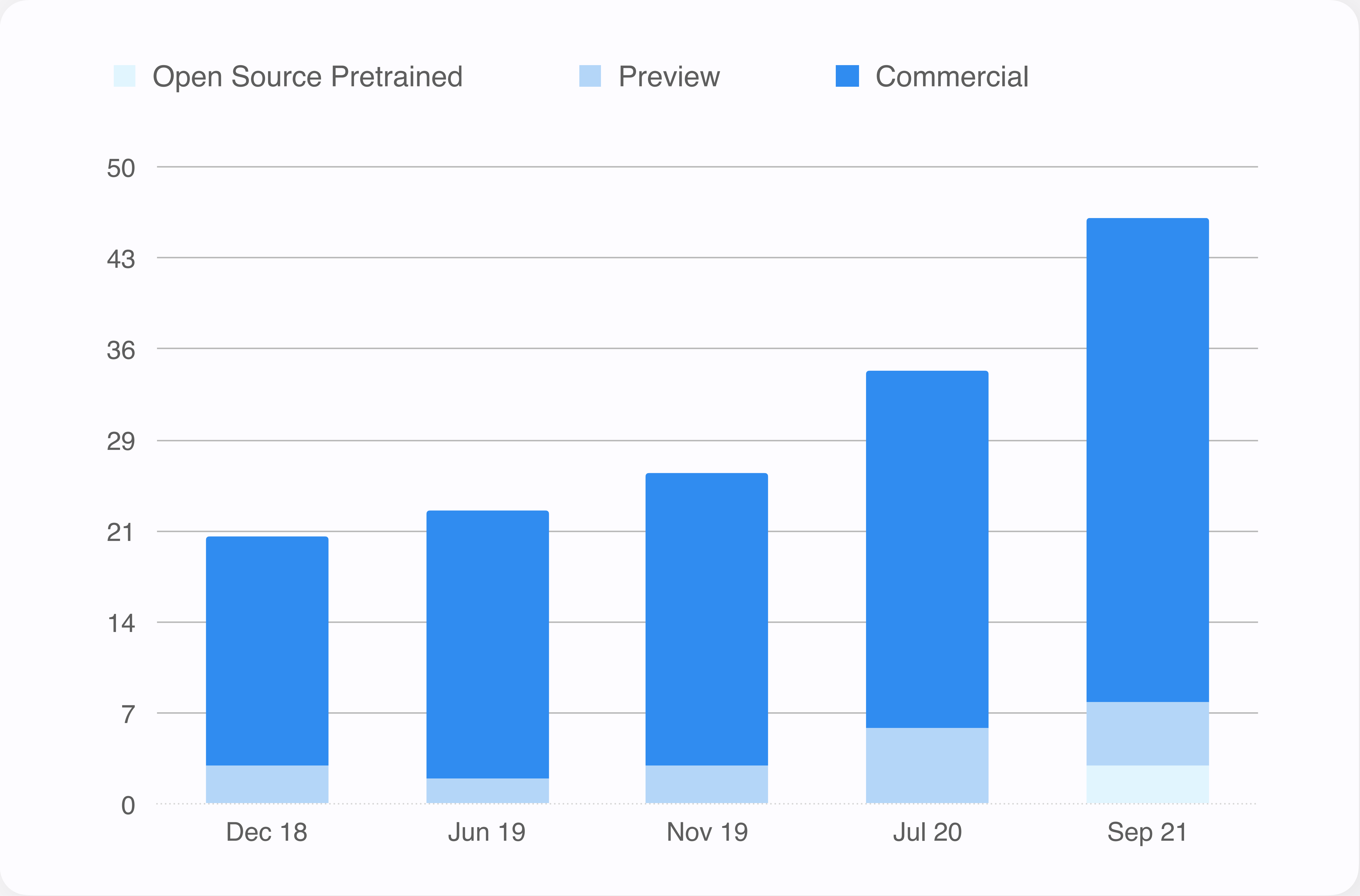
4.2 Public Pricing

USD per 1M symbols***

characters per month *	AISA	AppTek	Cloud Translation	GTCOM	Pangeanic	Prompsit	Rozetta	RWS Language Weather	XL8	SAP Translation Hub	Kawamura (NICT)	Globalese - v4	Alibaba Cloud	DeepL	IBM Watson	Google Translate	Naver Papago	ModernMT Realtime	Amazon	Systran PNMT	Microsoft	Tencent	Baidu	Youdao	Yandex Cloud	PROMT**	Niutrans	Elia	Tilde	Kakao
	on request									USD per 1,000,000 symbols																		USD per 1,000 words		free / beta
0										450	400	120	33	25	21	20	18	15	15	10	10	9	8	7	6	6	8	8	6	
500K											133																			
1M													15																	
3M											100																			
8M																														
10M																											6			
30M																														
32M																										5				
50M																											5			
64M																														
100M																						8								
128M												on request														4				
200M											on request																			
250M																					8									
500M																														
1B																					6					3				
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 Independent Cloud MT Vendors with Stock Models



Commercial (38)

AISA, Alibaba, Amazon, Apptek, Baidu, CloudTranslation, DeepL, Elia, Fujitsu, Globalese, Google, GTCom, IBM, iFlyTec, Lesan, Lindat, Lingvanex, Kawamura / NICT, Kingsoft, Microsoft, Mirai, ModernMT, Naver, Niutrans, NTT, Omniscien, Pangeanic, Prompsit, PROMT, Process9, Rozetta, RWS, SAP, Sogou, Systran, Tencent, Tilde, Viscomtec, Yandex, YarakuZen, Youdao

Preview / Limited (5)

eBay, Kakao, QCRI, Tarjama, Birch.AI

Open Source Pretrained (3)

M2M-100, mBART, OPUS

4.4 Open Source Pre-Trained MT Engines

OPUS MT
paper + code

The model was trained on openly available parallel corpora collected in the large bitext repository OPUS (Tiedemann, 2012). The architecture is based on a standard transformer setup with 6 self-attentive layers in both, the encoder and decoder network with 8 attention heads in each layer. Model is created by Language Technology Research Group at the University of Helsinki. It's based on Marian-NMT. And it is technically a separate model for each pair.

M2M-100
paper + code

The model was trained on 7.5B parallel sentences, corresponding to 2200 directions. The data was mined via CCMatrix and CCAligned. Transformer based architecture with 12 encoder and 12 decoder layers, with embedding dimension of 1024. We tested M2M-100 418M params and M2M-100 1.2B params. There is also a 12B-params Model, as we took models that can be used on a single GPU. Model is created by Facebook Research. The model is interesting in that it's not English-centric and can translate directly between any pair of 100 languages; a single model for '*' → '*'.

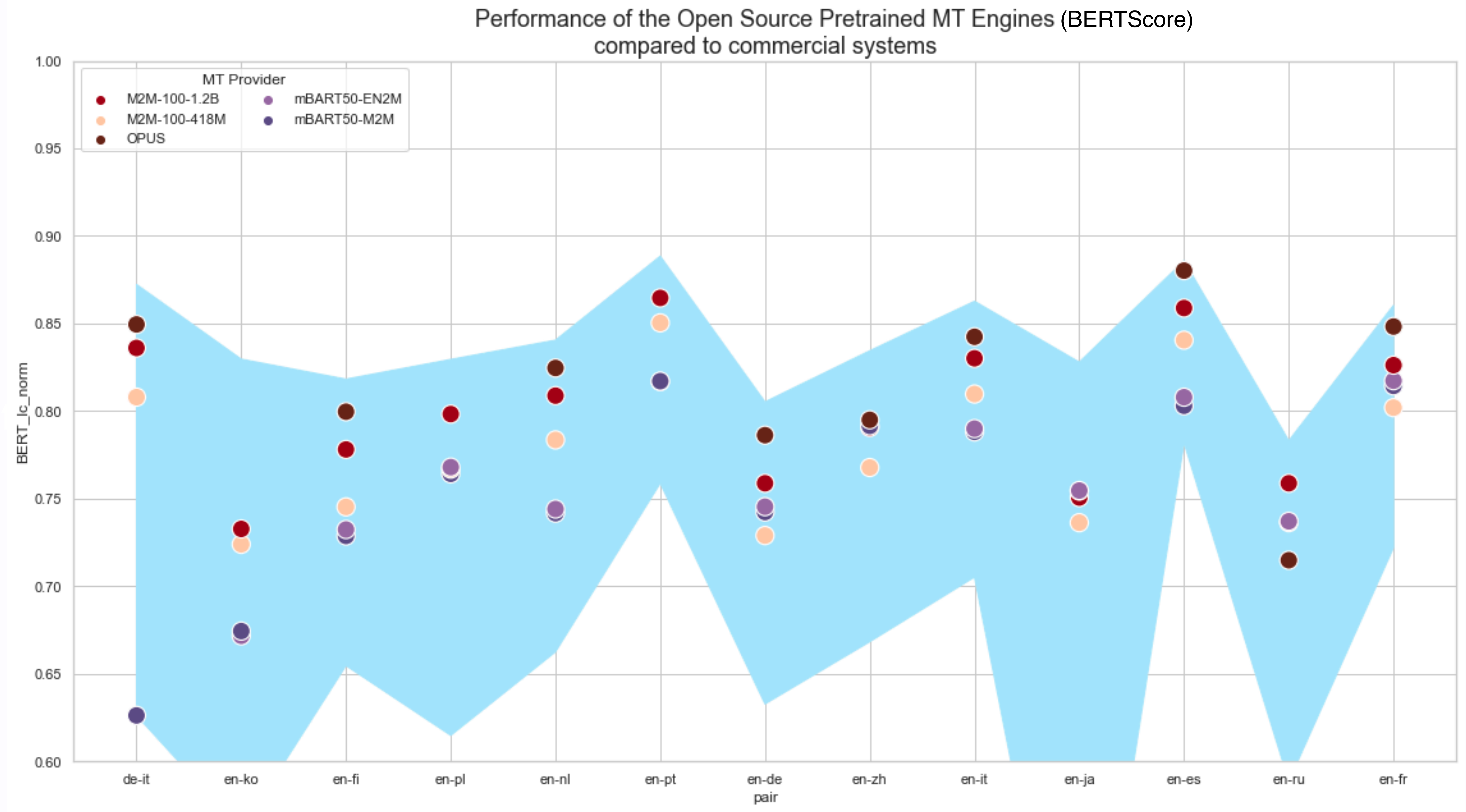
mBART50
paper + code

The model was trained on various sources for 50 languages including parallel and monolingual data (Common Crawl). Transformer based architecture with 12 layers of encoder and 12 layers of decoder with embedding dimension of 1024 and 16 heads. There is about ~ 680M parameters. Model was done by Facebook Research. mBART50 is based on mBART25. We used mMBART50 fine-tuned for two configurations one-to-many (a single model for en →*) and mMBART50 many-to-many a single multilingual model '*' → '*'.

Intento

Open Source MT Performance (BERTScore)

- **OPUS** and **M2M-100** mostly show performance in the 2nd tier of commercial systems.
- For en-es, **OPUS** scores are on par with the best commercial systems
- For **en-ko** and **en-ja**, the scores are very poor.
- **OPUS** leads for de-it, en-fi, en-nl, en-de, en-zh, en-it, en-es, and en-fr.
- **M2M-100** leads for en-ko, en-pl, en-pt, en-ru



5 Key Conclusions

The **MT market is accelerating. 13 more vendors** offer pre-trained MT models since August 2020, plus there are **three open-source** pretrained MT engines available. We have evaluated **29 MT engines — 14 more than a year ago!**

Unprecedented language coverage: **99,760 language pairs** across all MT engines. It was just 16K a year ago! The main contributors are **Niutrans** with their 88K language pairs and **Alibaba** with 20K.

19 MT engines are among the statistically significant leaders for **7** industry sectors and **13** language pairs. **9** MT engines provide minimal coverage for all language pairs and industries, **1-4** per industry sector.

Many engines show best results for English to **Spanish**, **Russian**, and **Chinese**. **Legal**, **Financial**, and **Healthcare** require a careful choice of MT vendor, as few perform at the top level. Despite having several comparable MT engines per language pair, **Education** shows relatively low scores, which may indicate the importance of customization in this domain.

Open-source engines perform in the 2nd tier of commercial systems, except for **en-es** (on par with top-tier systems) and **en-ko** & **en-ja** (much worse than commercial systems).

Intento Enterprise MT Hub

A gateway enabling global companies to unlock the full potential of AI for creative processes, **increasing productivity by 20X.**

Localization

Supercharge your Translation Management System with additional machine translation options through integration with MT Hub for Localization. Automatically select best-fit MT engines, keep track of MT usage, fine-tune MT output, and more.

Customer Service

Make your customers happy by providing exceptional multilingual customer support with no extra effort or headcount. Grow revenue, saving time and budget on global business expansion. Works in Zendesk, ServiceNow, Salesforce and more.

Office Productivity

Seamlessly translate any file in 190 languages – Word, Excel, PowerPoint, and PDF – while replicating the format of the original document. There's no need to reproduce anything manually.

Community & Marketing

Help your international customers decide to buy your products and services by providing access to product descriptions, reviews, and community discussions in their language in real-time. Works in Telligent, Jive, ServiceNow and other portals.

Software Development

Ensure flawless multilingual communication between development teams. Save their time translating Confluence documents and pages, Jira tickets, and even code reviews, comments, and discussions.

MT Studio

Train, evaluate, and improve the best-fit machine translation model from a single interface.

To know more go to https://inten.to

Book a live demo

Intento Plugins and Connectors

Localization

- MT Hub for Fluency Now
- MT Hub for Lingotek
- MT Hub for Matecat
- MT Hub for memoQ
- MT Hub for Memsource
- MT Hub for Trados
- MT Hub for Smartcat
- MT Hub for Smartling
- MT Hub for Wordfast
- MT Hub for Wordbee
- MT Hub for XTM Cloud

Office Productivity

- Translator for Word
- Translator for Excel
- Translator for Outlook
- Translator for Windows
- Translator for Mac
- Translator for Chrome
- Translation Portal

Software Development

- Translator for Chrome (any SaaS)
- Translator for Windows (any desktop app)
- Translator for Mac (any desktop app)
- Translator for Jira

Customer Service

- Translator for Zendesk Agents
- Translator for ServiceNow
- Translator for Chrome (works for any livechat: Salesforce, Intercom, Oracle and others)

Community & Marketing

- Translator for Telligent
- Translation in other community portals and KBs via frontend integration

To know more go to https://inten.to

Book a live demo

MT Evaluation with Intento

End-to-End

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

Fast and Safe

5–6 weeks from assorted TMs and glossaries to winning MT engines with effort saving estimation for Post-Editing Machine Translation and quality estimation for Real-time cases, such as support chats.

Trusted

We run 15–20 MT Evaluation projects per month for global companies across industries under strict Security, Quality and Data Protection requirements. ISO 27001 and ISO 9001 certified.

Reach us at hello@inten.to

The State of Machine Translation

Stock MT Models

Commercially available pre-trained MT models

data provided by TAUS the language data network

Intento, Inc.
hello@inten.to

2261 Market St, #4273
San Francisco, CA 94114

Intento

Independent Multi-Domain Evaluation of Machine Translation Engines

Part 2: Deep-dive linguistic analysis

In partnership with

[creative words] effectiff EURO-COM INTERNATIONAL logrus IT Multilingual Connections arcadia TRANSLATIONS

- 3 language pairs (EN → ES, EN → IT, EN → NL)
- Comparison of texts between 5 industry sectors: Education, Financial, Healthcare, Legal, Travel (ES)
- LQA results and the nuances indicated during the LQA phase
- Key conclusions on how automatic metrics relate to human estimation of translations
- Recommendations on the best-fit MT engines for analyzed language pairs and industry sectors
- Insights on how to enhance the power of all MT engines available on the market

Coming soon.

You will receive a link to Part 2 to the same email that you provided to get Part 1.

Appendix A.

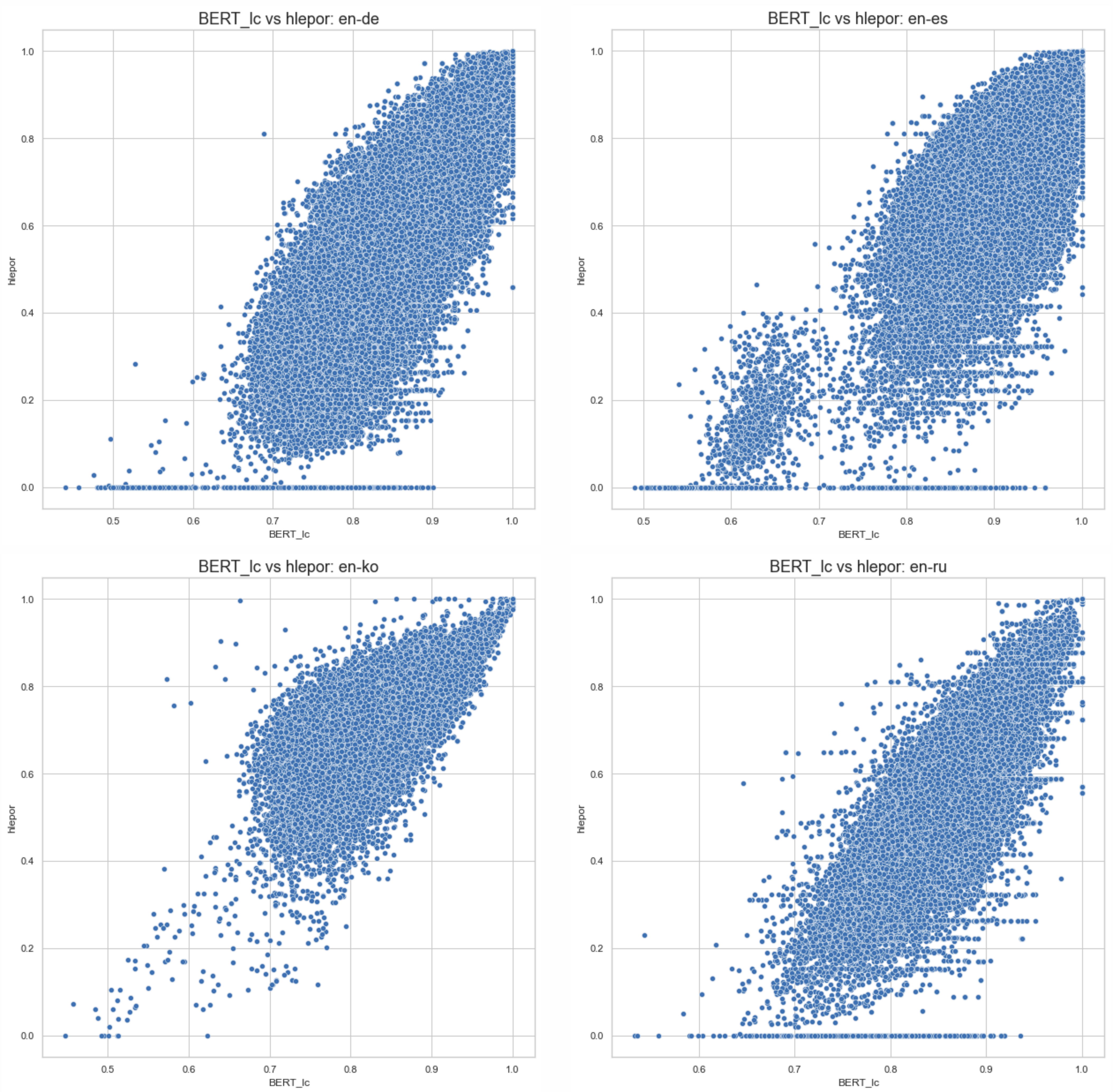
Comparing hLEPOR and BERTScore

low hLEPOR + high BERTScore

- paraphrases / synonyms
- minor punctuation / tokenization issues

high hLEPOR + low BERTScore

- mostly doesn't exist
- punctuation and spacing issues in Asian languages (our tokenization for hLEPOR doesn't penalize them)



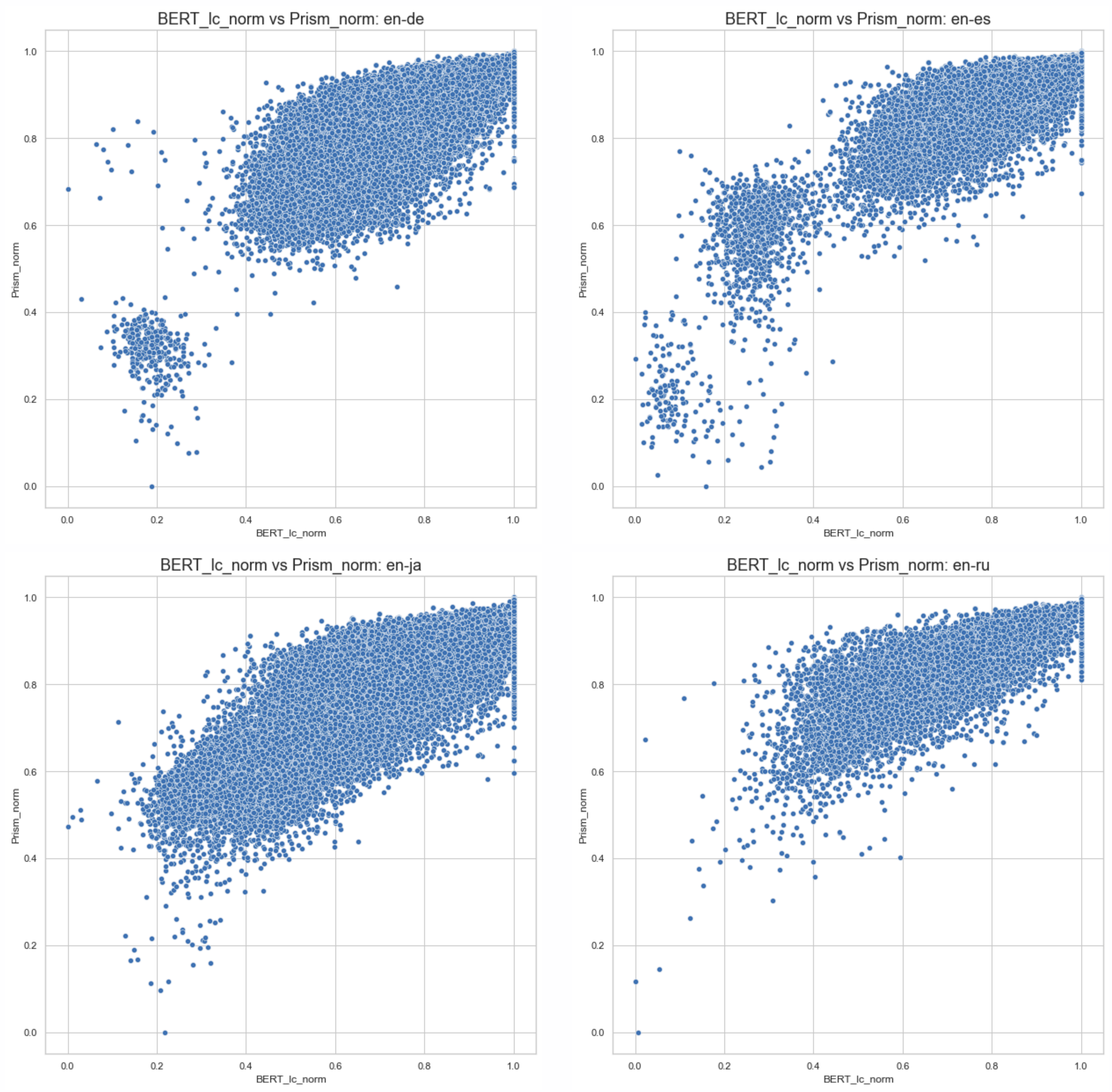
Comparing
BERTScore and PRISM

low BERTScore + high PRISM

- context-dependent alternative translations with different meanings (non-paraphrases)
- non-translated phrases

high BERTScore + low PRISM

- PRISM for identical translations is not guaranteed to be close to 1
- different capitalisation



Intento

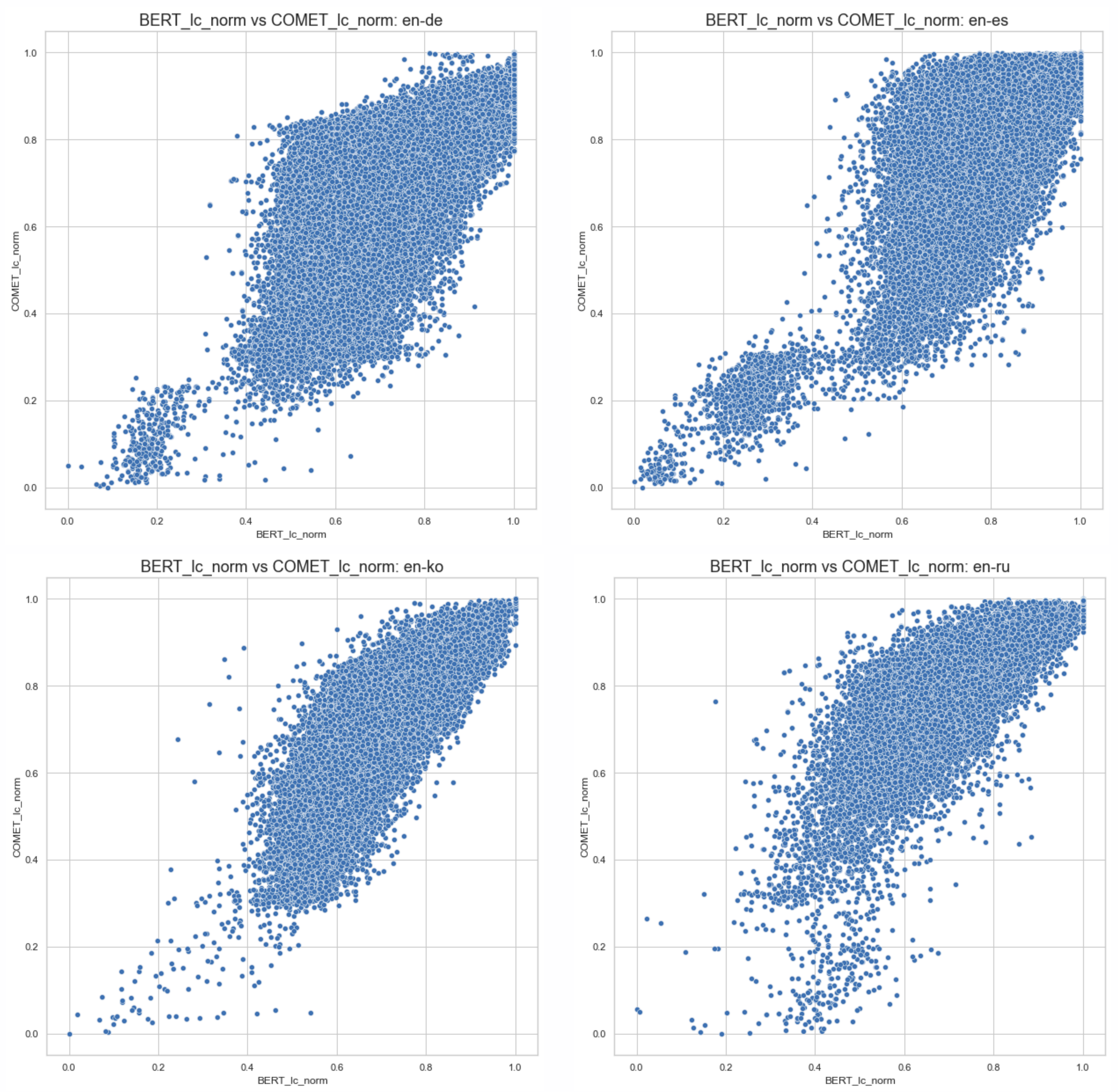
Comparing
BERTScore and COMET

low BERTScore + high COMET

- context-dependent alternative translations with different meanings (non-paraphrases)
- minor tokenization issues (e.g. merging words vs using "-" in German)

high BERTScore + low COMET

- omissions and omissive paraphrases
- context-dependent alternative translations with a different gender or tone of voice (mostly short sentences that lack context)



Intento

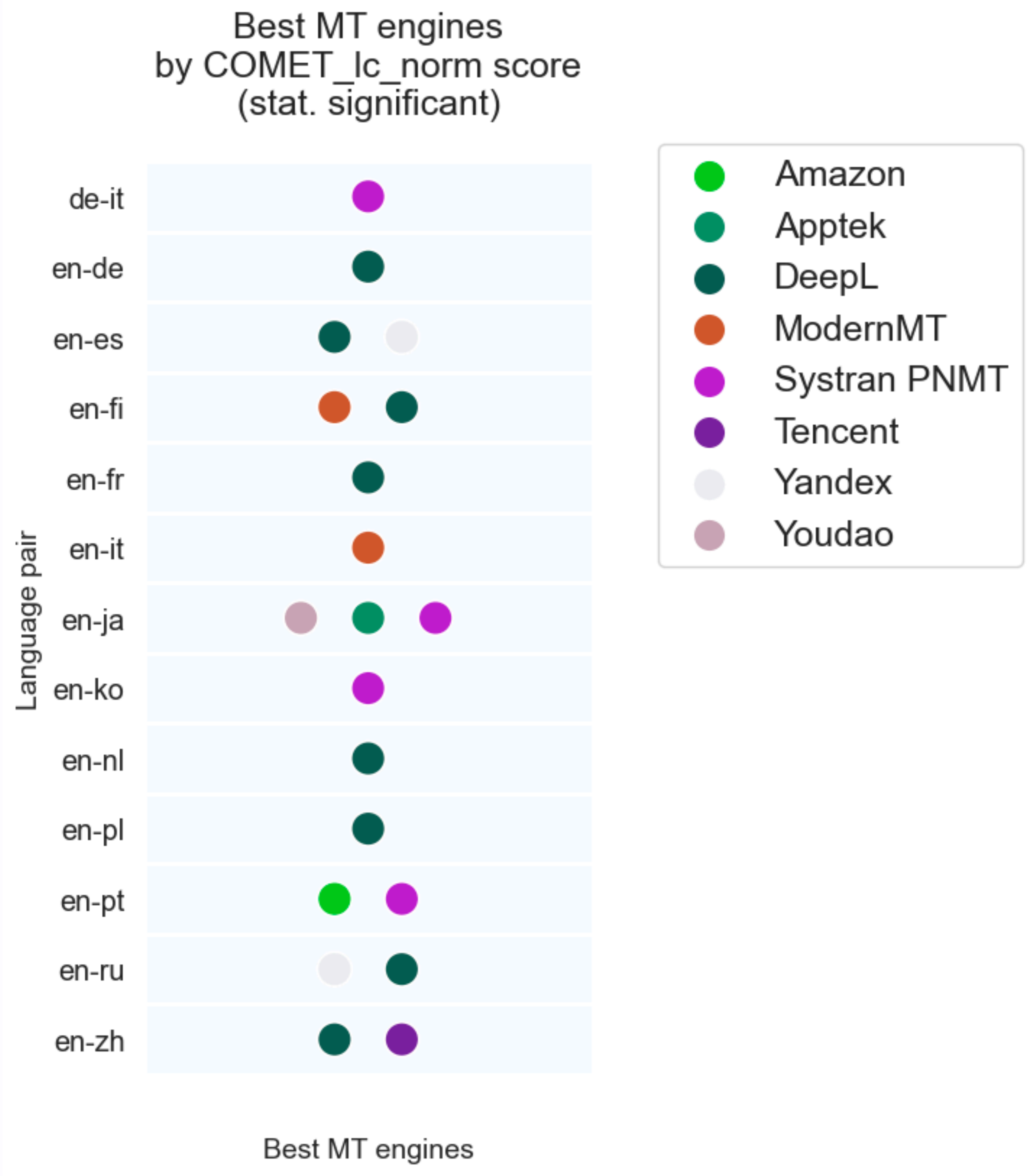
Appendix B.

A.1 Ranking for COMET Score

- Predicts machine translation quality using information from both the source input and the reference translation. Achieves state-of-the-art levels of correlation with human judgement.
- For every language pair, we have normalized COMET to fit [0,1] interval.
- COMET significantly penalizes different capitalization, therefore we have lowercased all text inputs. Per our observations, it does not lead to score corruption for properly capitalized sentences.
- Also, COMET seem to penalize different gender and other context-dependent factors, which mostly affects translations of short sentences with little embedded context.
- Hence, **we recommend using COMET to evaluate MT on your own data with the goal of choosing the engine closest in phrasing and ambiguity resolution to the reference.**
- Does not reflect absolute quality level. Not comparable across language pairs.
- In our research we used the unbabel-comet version 0.1.0 with the model for evaluation with reference translations - wmt-large-da-estimator-1719
- We are grateful to Unbabel for releasing the COMET metric and appreciate Unbabel's support and guidance in configuring it.

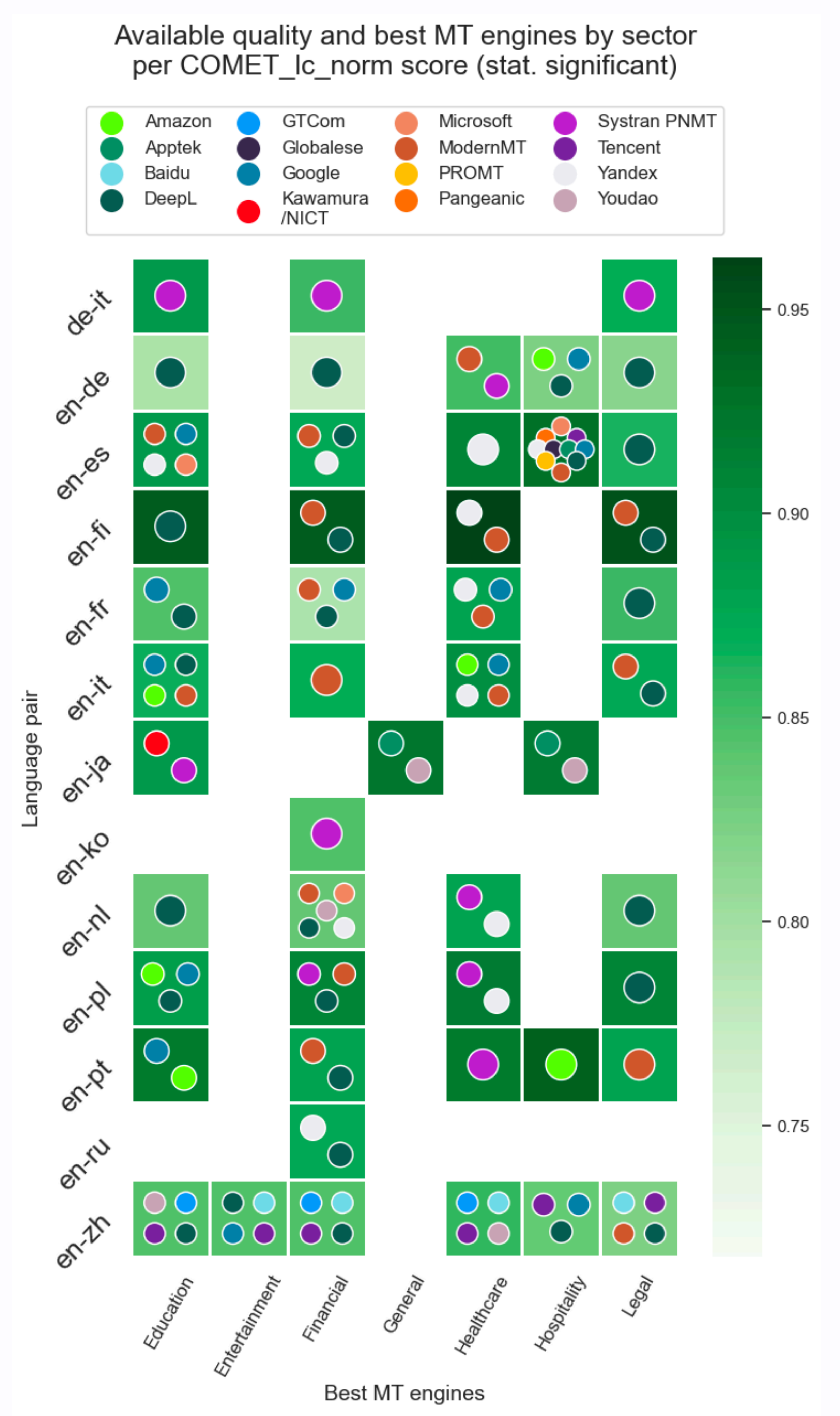
A.2 Best MT per Language Pair (COMET)

- Being more restrictive to alternative translations, there's just 8 leading MT engines, with much less per language pair.
- Fewer for minimal coverage: DeepL, Systran, and ModernMT.
- Absolute values are not shown to avoid confusion, as the scores are not comparable across language pairs.
- The domain and content type mix is different for every language pair (see the next slide) and greatly influences this leaderboard.



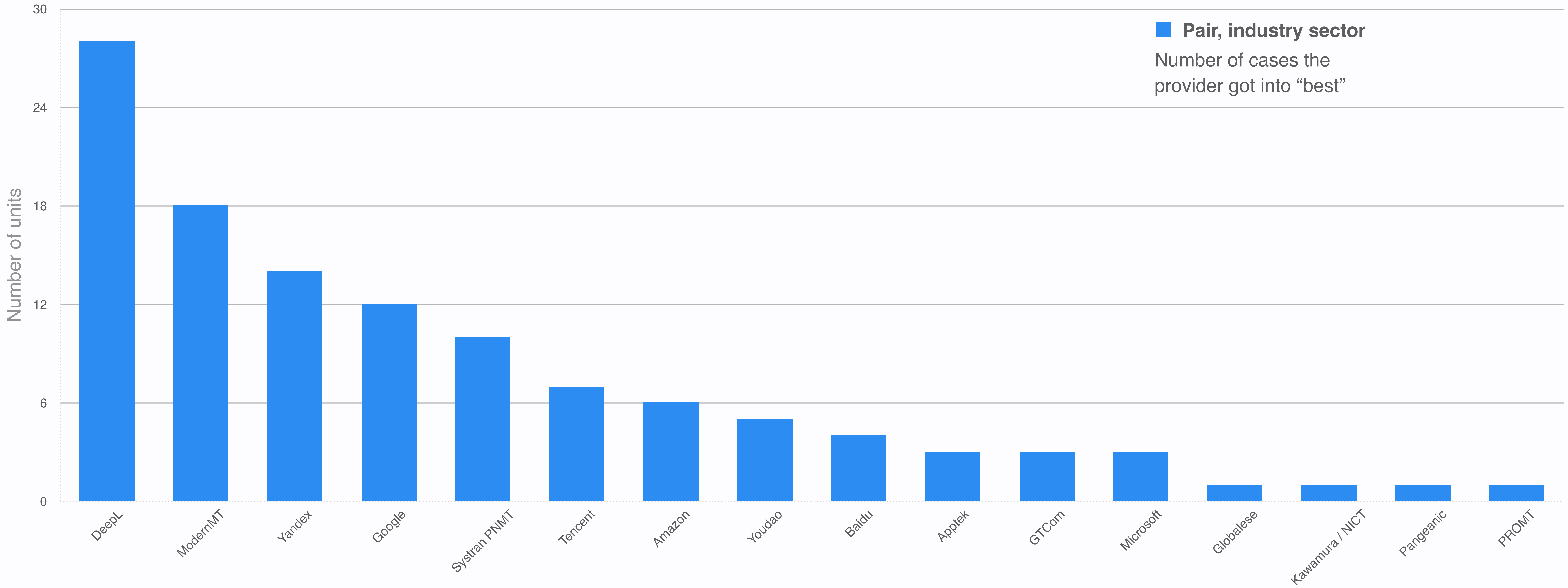
A.3 Best MT per Industry Sector (COMET)

- This chart is provided for reference. We recommend using BERTScore chart on Slide 23.
- 16 MT engines are among the statistically significant leaders for 7 industry sectors and 13 language pairs.
- The only significant difference from BERTScore is English to Portuguese, Financial domain, where leading MT is totally different for COMET.
- COMET may be appropriate when applied for post-editing, as per the score specifics.
- COMET favors DeepL a lot, our hypothesis - because DeepL is consistent in tone of voice and other context-dependent features, and resolves them similarly to the test set.



A.4 TOP Performing MT Providers (COMET)

Across 13 language pairs, 7 industry sectors



Appendix C.

C.1 Ranking for PRISM Score

B.1 Ranking for PRISM Score

- Evaluates machine translation as a paraphrase of a human reference translation. Penalizes both fluency and adequacy errors.
- For every language pair, we have normalized PRISM to fit [0,1] interval.
- PRISM penalizes different capitalization, but it also penalizes making texts lowercase, hence for PRISM we have decided to keep the capitalization as is.
- May not reach [0,1] for identical sentences, which makes its problematic to average across segments and draw conclusions for high-performing MT.
- Not available for Korean.
- **Because of the issues listed above, we do not provide ranking for PRISM to avoid confusion.**
- Does not reflect absolute quality level. Not comparable across language pairs.

Appendix D.

D.1 Scores for Sentences of Different Lengths

Intento

C.1 Scores for Sentences of Different Lengths

- Typically, the scores are higher for shorter sentences. The exceptions are en-fi, en-it, en-nl, en-pt
- English-to-Japanese demonstrates significant difference among MT engines for short and long segments (see the picture)
- Some MT engines provide the top-tier scores for short and medium sentences, but fail to translate long ones, leading to the low average performance:
  - Tencent for de-it
  - Amazon for en-de
  - PROMT for en-es
  - Google for en-it
  - Microsoft for en-pl