

# The State of Machine Translation 2022

An independent evaluation of MT engines

Konstantin Savenkov, CEO at Intento (speaker)

Michel Lopez, CEO at e2f

**31** MT Engines

**11** Language pairs

**9** Industry sectors

# Agenda

1. Datasets
2. Evaluation methodology
3. Evaluation results
4. Miscellaneous
5. Key conclusions

GET FULL REPORT AT  
<https://bit.ly/mt-2022>

31

Machine Translation Engines

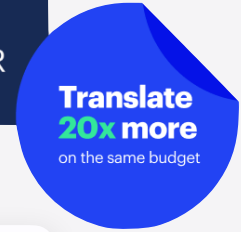
11

Language Pairs

9

Industry sectors

# About Intento



Intento allows global enterprises to translate 20x more on the same budget. It helps evaluate, select, customize, and connect best-fit AI with existing software and vendors. With Intento, businesses can also monitor translation performance to continuously improve their entire machine translation program.

Trusted by Global Enterprise



# About e2f

Established in 2004, e2f helps people and machines understand each other fluently, regardless of language, content, and culture. e2f solutions empower Fortune 50 brands to monitor, objectively assess, and improve communications on a global scale.

e2f delivers world-class translation and training data with its proprietary technology stack for translation, quality review, and AI services. e2f offers a global resource pool of skilled professionals in virtually all countries and languages.

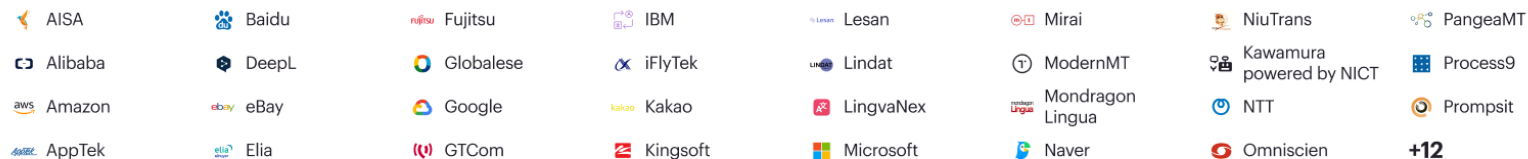
To learn more, [contact e2f](#) or [visit website](#).

## e2f services

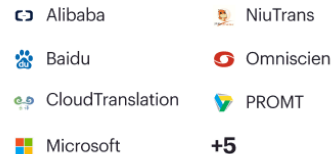
- MT detection and MT quality evaluation services that enable organizations to monitor suppliers for compliance with brand standards for human and machine translation.
- Creation of custom Lingosets™, or augmented multilingual datasets that represent real human conversational flow. Lingosets serve as benchmarks for conversational AI deployments.
- Golden datasets and training datasets that enable leading MT providers to evaluate and fine-tune engine performance.

# Machine Translation Landscape

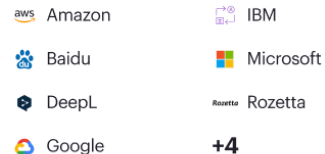
## Generic stock models



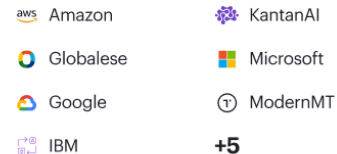
## Vertical Stock Models



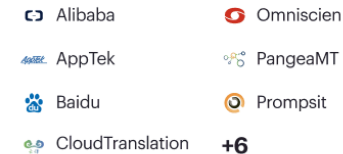
## Custom terminology support



## Auto domain adaptation































## Manual domain adaptation



# Machine Translation Engines

Evaluated in the study

Customization options:  none  TM  glossary  both

 AISA Neural Machine Translation API <input type="radio"/>	 Alibaba eCommerce MT <input type="radio"/>	 Alibaba Cloud General <input type="radio"/>	 Amazon Translate <input checked="" type="radio"/>	 Apptek Neural Machine Translation <input type="radio"/>
 Baidu Translate API <input checked="" type="radio"/>	 DeepL API <input checked="" type="radio"/>	 Elia Elhuyarren itzultzaile automatikoa <input type="radio"/>	 Globalese Machine Translation <input type="radio"/>	 Google Cloud Advanced Translation <input checked="" type="radio"/>
 GTCOM YeeCloud MT <input type="radio"/>	 IBM Watson eCommerce MT <input checked="" type="radio"/>	 Meta AI NLLB <input type="radio"/>	 Microsoft Language Translator <input type="radio"/>	 ModernMT Realtime <input checked="" type="radio"/>
 Naver Papago NMT Commercial <input type="radio"/>	 NiuTrans Translation Cloud Platform <input type="radio"/>	 Pangeanic Machine Translation API <input type="radio"/>	 PROMT Cloud API <input type="radio"/>	 RoyalFlush Finance Translation <input type="radio"/>
 Rozetta T-400 Machine Translation API <input checked="" type="radio"/>	 SYSTRAN PNMT <input checked="" type="radio"/>	 Tilde Machine Translation API <input type="radio"/>	 Tencent Cloud TMT API <input type="radio"/>	 Ubiquis Translation API <input checked="" type="radio"/>
 Yandex Translate API <input checked="" type="radio"/>	 Youdao Cloud Translation API <input type="radio"/>	 XL8 Machine Translation <input type="radio"/>		

# Datasets — Preparation

## Translation

- Selected native translators with expert-level qualifications and positive feedback in each language and domain.
- For reviews, selected native language experts in editing and proofreading across multiple domains, and positive customer feedback.
- Proofread strings supplied by Intento for compliance with proper English grammar, spelling, and punctuation and supplied files to translators via e2f's Translation, Editing, and Proofreading (TEP) platform.

To mitigate the possibility that a supplier could gain an unfair advantage by training an engine against the same dataset used for evaluation, Intento commissioned e2f to build an original golden dataset for this year's study.

## Quality Assurance

Provided via e2f's TEP portal

- Human translations were compared with ones generated by the leading machine translation engines using e2f's MT Detection tool, and accessed the probability that they contained machine-translated and/or post-edited content (MTPE).
- Strings whose MTPE probability exceeded e2f's threshold triggered expert review and was followed by re-translations, which were automatically reassessed. **The resulting golden dataset does not bear traces of MTPE.**
- Quality assurance reports were run on capitalization, punctuation, spelling, numbers, spaces, and typos. Reviewers implemented necessary changes and proofread the dataset prior to final delivery.

# Datasets — Preparation

- **9** industry sectors per language pair
- **500** segments in **11** language pairs per industry sector
- This year, we have identical segment coverage for all language pairs.





# Content Samples

## Industry Sectors

### General

*"Walmart is also the largest grocery retailer in the United States."*

### Finance

*"Both operating profit and net sales for the three-month period increased, respectively from €16m and €139m, as compared to the corresponding quarter in 2006."*

### Hospitality

*"Very reasonably priced and the food is excellent, I had pasta which was delicious, and my friend had the Italian meats & cheeses."*

### Healthcare

*"Leishmaniosis caused by Leishmania infantum is a parasitic disease of people and animals transmitted by sand fly vectors."*

### Legal

*"Landlord and Tenant acknowledge and agree that the terms of this Amendment and the Existing Lease are confidential and constitute proprietary information of Landlord and Tenant."*

### Entertainment

*"Further, they are aided by a magnificent cast of co-stars, most notably their secretary, played by Isabel Tuengerthal, who is a rare gem with great comic potential."*

### Education

*"Find what straight lines are represented by the following equation and determine the angles between them."*

### IT

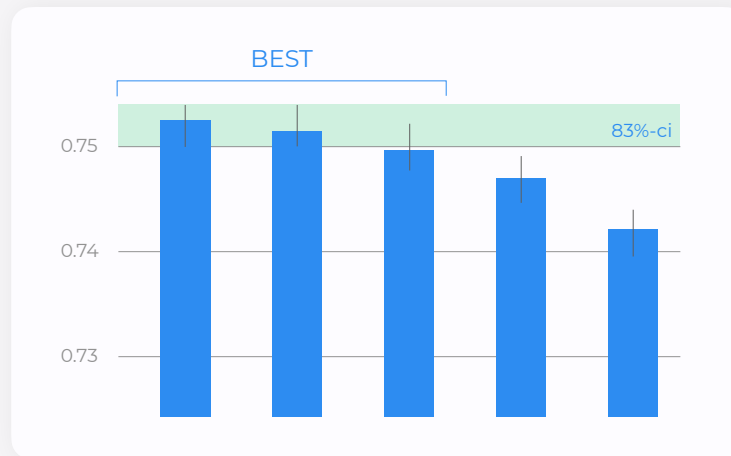
*"Result shows that GPU based the stream processor architecture ate more applicable to some related applications about neural networks than CPU."*

### Colloquial

*"and, in fact, there are two huge lenses that frame the figure on either side".*

# 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<sup>1,2</sup>



<sup>1</sup> 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.

<sup>2</sup> 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

# 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.



# 122,831 Language Pairs Across All MT engines\*



From 99,760 in August'20 to 122,831 in August'21

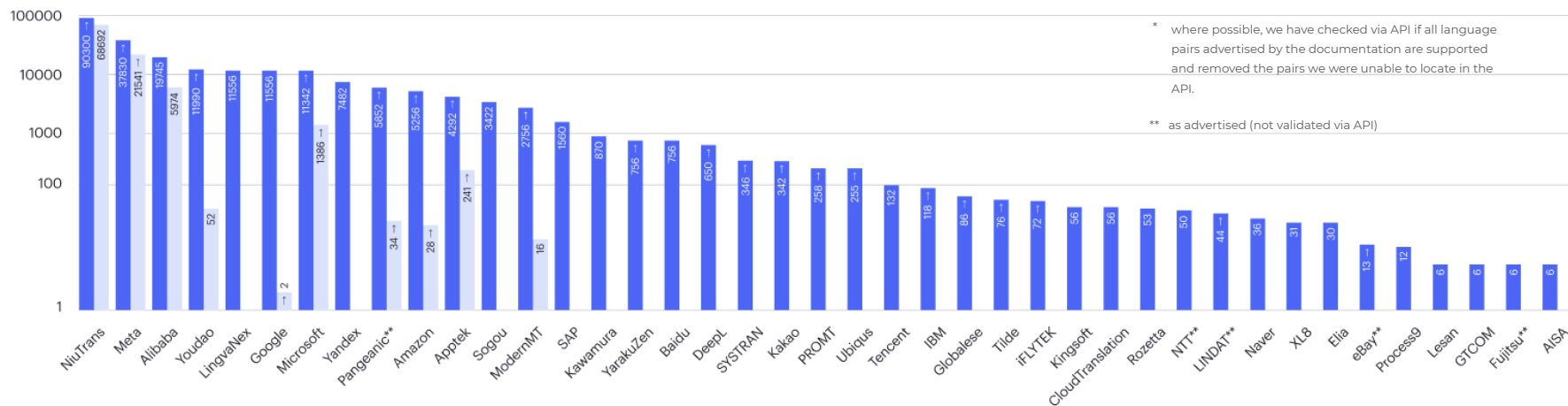


Significant growth for Microsoft, ModernMT and Amazon

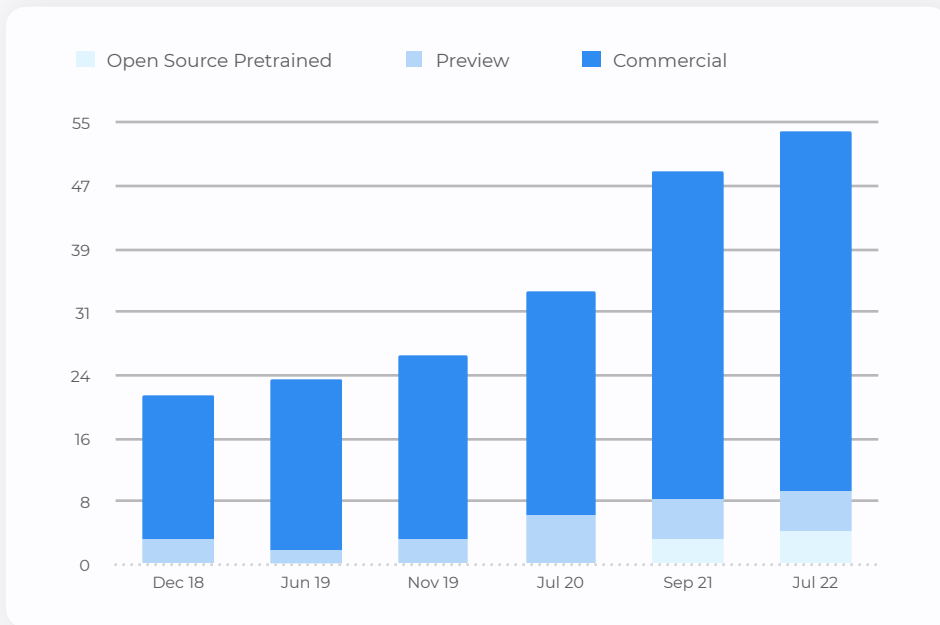


Added new niche MT providers with few languages.

● total language pairs ● unique language pairs



# Independent Cloud MT Vendors with Stock Models



## Commercial (45)

AISA, Alibaba, Amazon, Apptek, Baidu, CloudTranslation, DeepL, Elia, Fujitsu, Globalese, Google, GTCOM, IBM, iFlyTec, [HiThink RoyalFlush](#), Lesan, Lindat, Lingvanex, Kawamura / NICT, Kingsoft, [Masakhane](#), Microsoft, Mirai, ModernMT, Naver, Niutrans, NTT, Omniscien, Pangeanic, Prompsit, PROMT, Process9, Rozetta, RWS, SAP, Sogou, Systran, Tencent, Tilde, [Ubiquis](#), Viscomtec, [XL8](#), Yandex, YarakuZen, Youdao

## Preview / Limited (5)

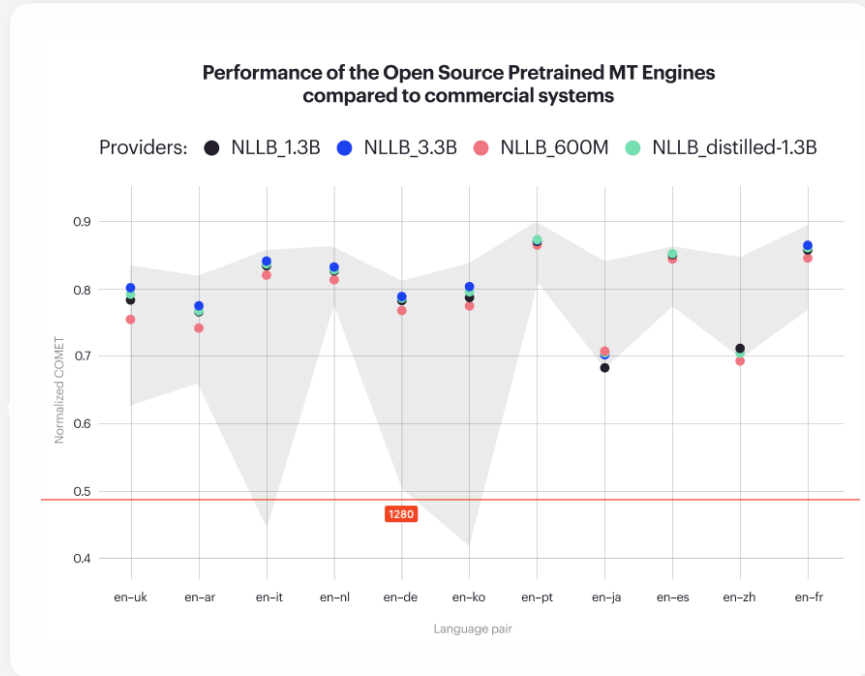
eBay, Kakao, QCRI, Tarjama, Birch.AI

## Open Source Pretrained (3)

M2M-100, mBART, [NLLB by Meta](#), OPUS

# Open Source MT Performance (BERTScore)

- **NLLB** by Meta AI mostly show performance in the 2nd tier of commercial systems.
- For en-es, **NLLB** scores are on par with the best commercial systems
- For **en-zh** and **en-ja**, the scores are quite low.
- **NLLB** with 3.3B parameters leads for en-uk, en-ar, en-it, en-nl, en-de, en-ko, and en-fr.
- **NLLB** with 1.3B parameters (distilled) leads for en-pt and en-es.



# Key takeaways



The **MT market is growing**. **4 more vendors** offer pre-trained MT models since August 2020, plus there are one new **open-source** pre-trained MT engine available (NLLB from Facebook). We have evaluated **31 MT engines - 2 more than a year ago**.



**Unprecedented language coverage: 122,831 language pairs** across all MT engines. It was 99K a year ago. The main contributors are **Niutrans** with their 90K language pairs, **NLLB by Meta** with 38K, and **Alibaba** with 20K.



**16** MT engines are among the statistically significant leaders for **9** industry sectors and **11** language pairs. **6** MT engines provide minimal coverage for all language pairs and industries, **2-4** per industry sector.



Many engines perform best with English to **Spanish** and **Chinese**. **Legal, Financial, IT, 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, **Entertainment** and **Colloquial** show 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-zh & en-ja** (much lower than commercial systems).



**New scores on the block!** This time, we have selected COMET as the main score based on the high correlation with human judgement in other evaluation projects.

GET FULL REPORT AT  
<https://bit.ly/mt-2022>



## BONUS TRACK 1: Score correlation with human judgement (based on another research of Intento)

We have run a separate study on 15 language pairs and 21 unique MT models, where we compared several metrics with human reviewers' judgement.

We found that in 10 out of 15 language pairs COMET has a better correlation with human ratings than other metrics, in 3 out of 15 language pairs BERTScore shows slightly better correlation, and in 2 language pairs based only on the data we currently possess both BERTScore and COMET show lower correlation results.

Please note that we have analyzed the post-editing case, and for other use cases, such as gisting or understanding MT, BERTScore may be better.

**Pearson correlation in en-de**

rating	1.00000	0.0423	0.0769	-0.0940	0.1585
BERTScore	0.0423	1.00000	0.7998	-0.7926	0.5894
hLEPOR	0.0769	0.7998	1.00000	-0.8921	0.4962
TER	-0.0940	-0.7926	-0.8921	1.00000	-0.5069
COMET	0.1585	0.5894	0.4962	-0.5069	1.00000

**Pearson correlation in en-pt**

rating	1.00000	0.0976	0.0684	-0.1191	0.1667
BERTScore	0.0976	1.00000	0.7840	-0.7709	0.5049
hLEPOR	0.0684	0.7840	1.00000	-0.9062	0.4256
TER	-0.1191	-0.7709	-0.9062	1.00000	-0.4276
COMET	0.1667	0.5049	0.4256	-0.4276	1.00000

**Pearson correlation in en-nl**

rating	1.00000	0.1482	0.1648	-0.1653	0.2881
BERTScore	0.1482	1.00000	0.8406	-0.8355	0.6019
hLEPOR	0.1648	0.8406	1.00000	-0.8876	0.4732
TER	-0.1653	-0.8355	-0.8876	1.00000	-0.5088
COMET	0.2881	0.6019	0.4732	-0.5088	1.00000

**Pearson correlation in en-fr**

rating	1.00000	0.1545	0.1463	-0.1838	0.2477
BERTScore	0.1545	1.00000	0.7897	-0.8421	0.6427
hLEPOR	0.1463	0.7897	1.00000	-0.8978	0.5995
TER	-0.1838	-0.8421	-0.8978	1.00000	-0.6158
COMET	0.2477	0.6427	0.5995	-0.6158	1.00000

**Pearson correlation in en-es**

rating	1.00000	0.0233	0.0202	-0.0258	0.1793
BERTScore	0.0233	1.00000	0.8233	-0.8315	0.4637
hLEPOR	0.0202	0.8233	1.00000	-0.9184	0.4570
TER	-0.0258	-0.8315	-0.9184	1.00000	-0.4499
COMET	0.1793	0.4637	0.4570	-0.4499	1.00000

**Pearson correlation in en-ko**

rating	1.00000	0.1742	0.1537	-0.0489	0.2721
BERTScore	0.1742	1.00000	0.8068	-0.8200	0.4488
hLEPOR	0.1537	0.8068	1.00000	-0.7890	0.4676
TER	-0.0489	-0.8200	-0.7890	1.00000	-0.4098
COMET	0.2721	0.4488	0.4676	-0.4098	1.00000

## BONUS TRACK 2: Classic BLEUs Hit

We present highest scores in each combination of sector and language pair.

The score here is solemnly corpus-based as BLEU does not provide segment scores due to its specifics.

Please keep in mind that BLEU, as a corpus-level score with a number of parameters, is not comparable not only across different languages but also across different datasets and different BLEU implementations.

Highest BLEU score for pair x domain

