



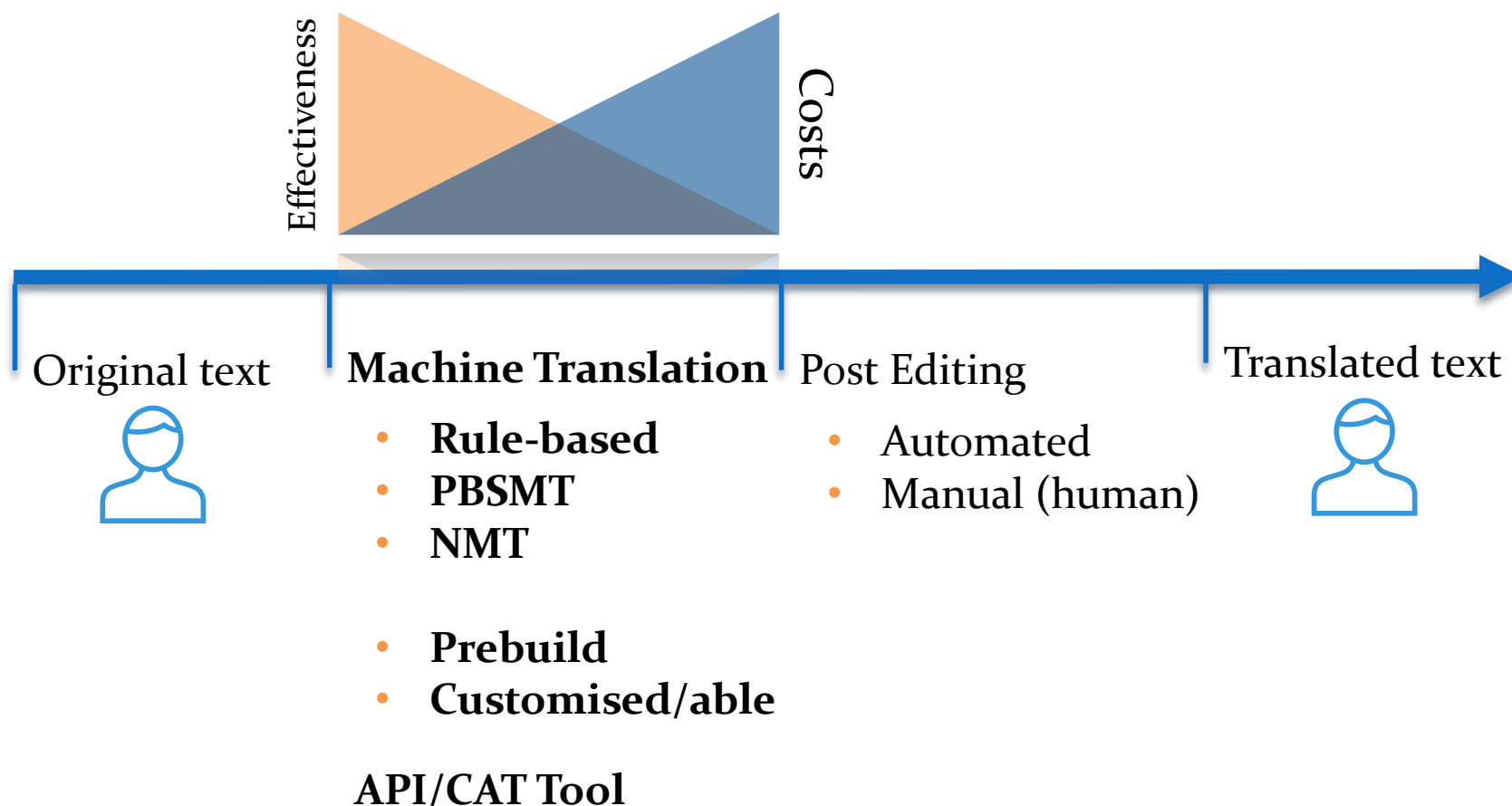
KantanMT.com
No Hardware. No Software. No Hassle MT.

Empirical evaluation of NMT and PBSMT quality for large-scale translation production.

*Dimitar Shterionov,
Pat Nagle,
Laura Casanellas,
Riccardo Superbo,
Tony O'Dowd*

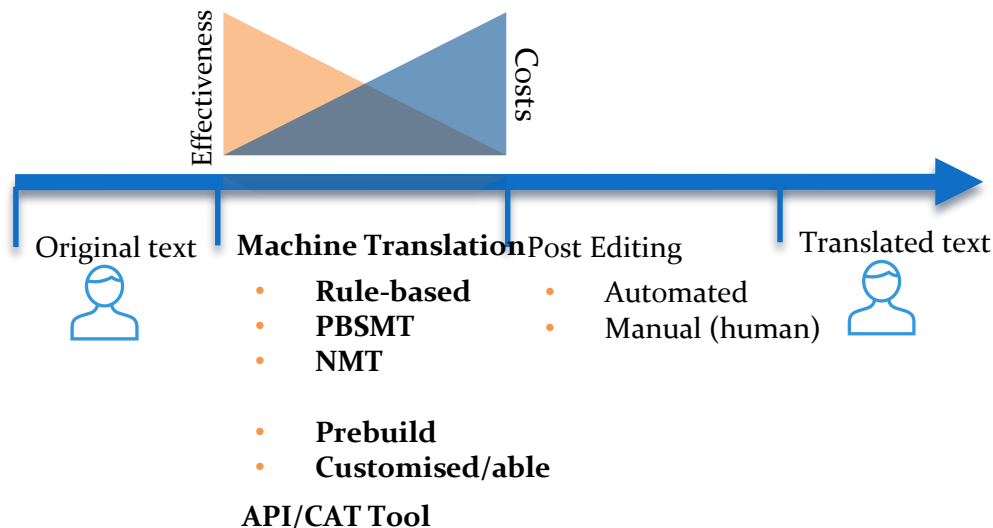
EAMT 2017, 29, May, 2017, Prague, the Czech Republic

MT-centric translation production line



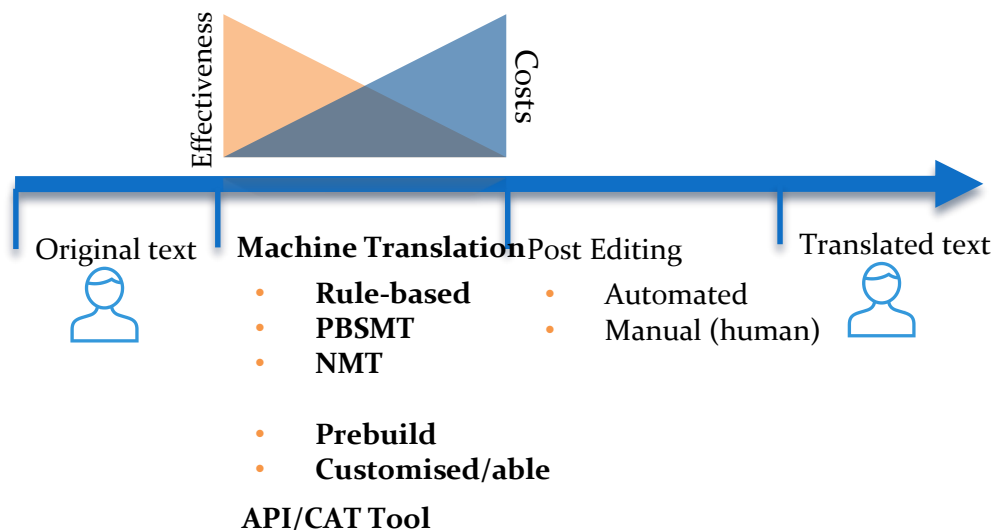
MT-centric translation production line

- **Can NMT be better than PBSMT?**



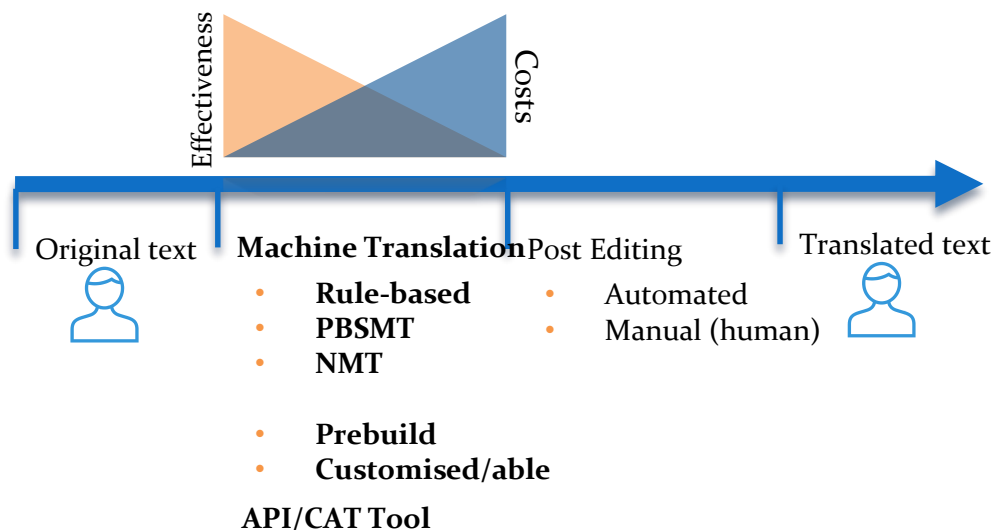
MT-centric translation production line

- **Can** NMT be better than PBSMT?
- How to evaluate and compare MT quality?

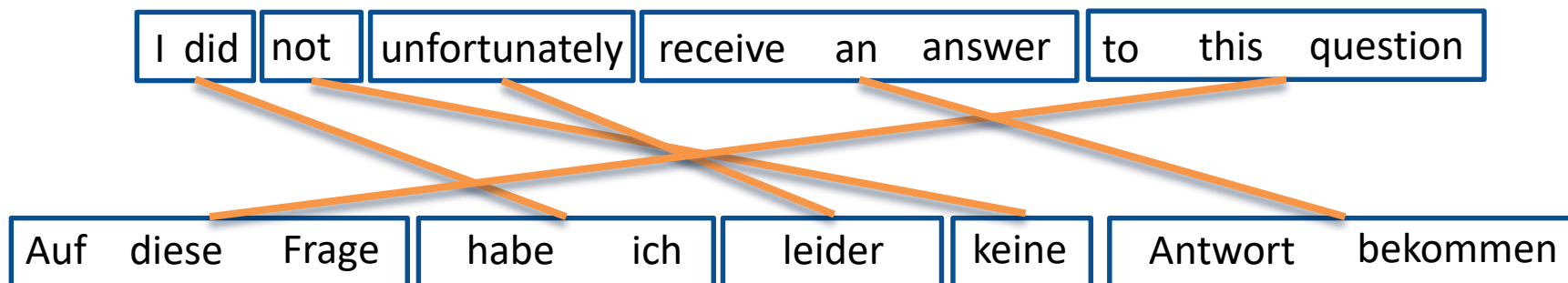


MT-centric translation production line

- **Can** NMT be better than PBSMT?
- How to evaluate and compare MT quality?
- Is NMT feasible for a large-scale translation production?



Phrase-based Statistical MT



- **Multiple components, sequentially connected**

- Translation model
- Language model
- Recasing model

- **Translation**

- A phrase translation is derived from the phrase table
- Language and recasing models add meaning

...

I did→hebe ich

I did→ich hebe

Unfortunately→leider

Unfortunately→unglücklich

Receive an asnwer→emfange eine Antwort

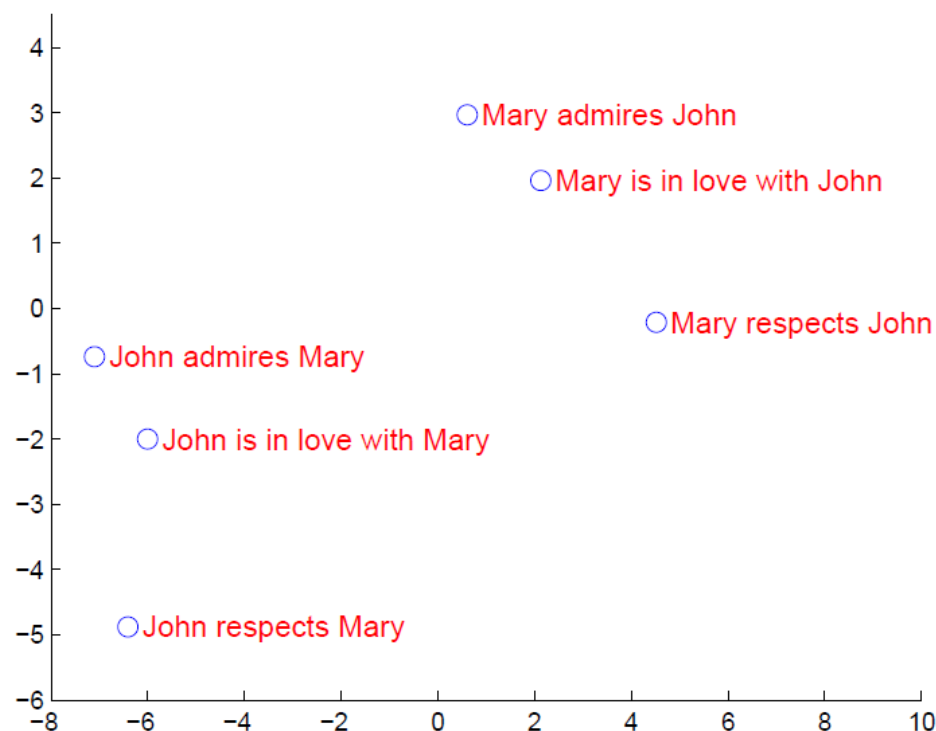
Receive an answer→Antwort bekommen

Receive an answer→Antwort erhalten

...

Neural MT

- **Encode-decoder neural network**
 - Two connected RNNs .
 - Trained simultaneously to maximise performance.
- **Training/Translation**
 - A source sentence is encoded (*summarised*) as a vector *c*.
 - Words segmented in word-units
 - The decoder predicts a word from *c* and already predicted words.



[Sutskever 2014] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, **Sequence to Sequence Learning with Neural Networks**

NMT vs. PBSMT

- **PBSMT considers phrases (1-grams ... n-grams); all phrases.**
- **NMT handles the sentence as a whole.**
- **PBSMT will translate each phrase or leave them untranslated.**
- **NMT will aim to translate everything;**
“unknown” will replace untranslatable.
- **PBSMT is more literal – can be more accurate.**
- **NMT can be more fluent – can be completely inaccurate.**
- **PBSMT is transparent -- easy to tamper with and improve.**
- **NMT is a “black box”.**

- **PBSMT and NMT are both data-driven MT paradigms.**

Empirical evaluation

- **Quality evaluation metrics**
 - BLEU
 - F-Measure
 - TER
- **Human evaluation:**
Side-by-side comparison

What is BLEU? (Papineni et al., 2002)

- Measures the precision of an MT system.
- Compares the n-grams ($n \in \{1..4\}$) of a candidate translation with those of the corresponding reference.
- The more n-gram matches the higher the score.
- Can be document- or sentence- level
- Factors for BLEU
 - Translation length
 - Translated words
 - Word order

[Papineni et al. 2002] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. **BLEU: A Method for Automatic Evaluation of Machine Translation**. In ACL 2002.

An example...

- **Source (EN):**

All dossiers must be individually analysed by the ministry responsible for the economy and scientific policy.

- **Translations (DE):**

1. Jeder Antrag wird von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik individuell geprüft.
2. Alle Unterlagen müssen einzeln analysiert werden von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik.
3. Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden.

An example...

- **Source (EN):**

All dossiers must be individually analysed by the ministry responsible for the economy and scientific policy.

- **Translations (DE):**

- | | | | |
|-------|-----------|----|---|
| | Reference | 1. | Jeder Antrag wird von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik individuell geprüft. |
| PBSMT | BLEU 58% | 2. | Alle Unterlagen müssen einzeln analysiert werden von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik. |
| NMT | BLEU 0% | 3. | Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden. |

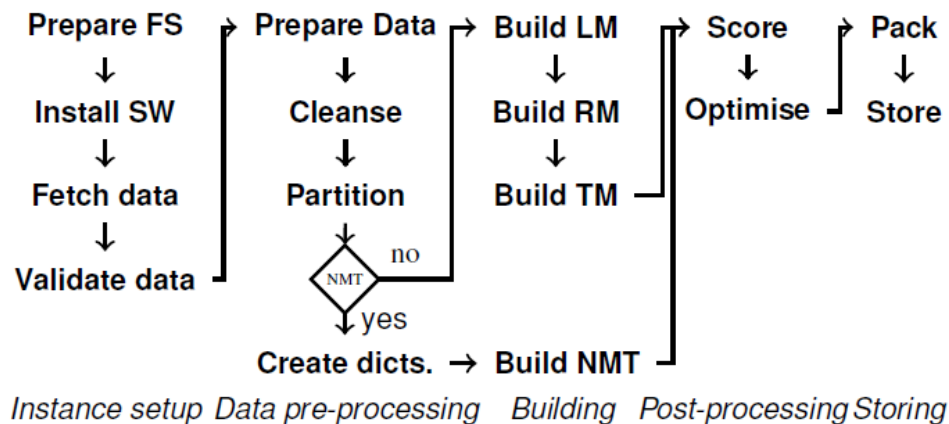
Empirical evaluation

- **Data:**

- EN-DE (8,820,562), EN-ES (3,681,332), EN-IT (2,756,185), EN-JA (8,545,366), EN-ZH-CN (6,522,064)
- Locked train, tune, test data

- **Systems:**

- PBSMT: Moses, CPU, FA, 5-gram LM, tuned 25 iter.
- NMT: OpenNMT, GPU NVIDIA K520, ADAM, 0.0005, batch: 64



- **Restrictions on the NMT training:**

- For no longer than 4 days
- Perplexity needs to be below 3

Empirical evaluation

- **Automatic quality evaluation:**
 - BLEU
 - F-Measure
 - TER

Empirical evaluation

- **Automatic quality evaluation:**

- BLEU
- F-Measure
- TER

	PBSMT				NMT				
Lang. pair	F-Measure	BLEU	TER	T	F-Measure	BLEU	TER	P	T
EN-DE	62	53.08	54.31	18	62.53	47.53	53.41	3.02	92
EN-ZH-CN	77.16	45.36	46.85	6	71.85	39.39	47.01	2	10
EN-JA	80.04	63.27	43.77	9	69.51	40.55	49.46	1.89	68
EN-IT	69.74	56.98	42.54	8	64.88	42	48.73	2.7	83
EN-ES	71.53	54.78	41.87	9	69.41	49.24	44.89	2.59	71

Empirical evaluation

- **Human evaluation:**

- Side-by-side (with KantanLQR / ABTesting)
- 200 sentence triples
- Native speakers of the target language; proficient in English

Empirical evaluation

- **Human evaluation:**

- Side-by-side (with KantanLQR / ABTesting)
- 200 sentence triples
- Native speakers of the target language; proficient in English

Add Test Data

1 Source Paste your source segments.

2 AB Target Paste your target segments.

3 CD Target Paste your optional target segments.

4 Summary Check details and add test data.

Source Segments

The same holds for its supporting scripts config.guess config.sub ltconfig ltmain.sh. Planned for availability in Q1 2008 EMC will be delivering a Thin Provisioning capability for Symmetrix DMX systems.

Previous Next Cancel Add

Add Test Data

1 Source Paste your source segments. ✓

2 AB Target Paste your target segments.

3 CD Target Paste your optional target segments.

4 Summary Check details and add test data.

Target A Segments

戴尔专业技术支持为最终用户提供小型企业所有者需要的 24 x 7 的“how-to”的援助可以针对部分常见软件应用程序，以及帮助他们搭建简单的无线和有线网络。独特的戴尔制造流程要求如果没有声音，请单击“注释”选项卡

Target B Segments

面向最终用户的 Dell ProSupport 可为小型企业所有者提供全天候所需的全天候“how-to”帮助，选择常见的软件应用程序，并帮助简单的无线和有线网络。独特的戴尔制造流程要求如果没有声音，请单击“注意”选项卡，指导您完成本课程。

Previous Next Cancel Add

Empirical evaluation

- **Human evaluation:**

- Side-by-side (with KantanLQR / ABTesting)
- 200 sentence triples
- Native speakers of the target language

Scored segments: 71/100

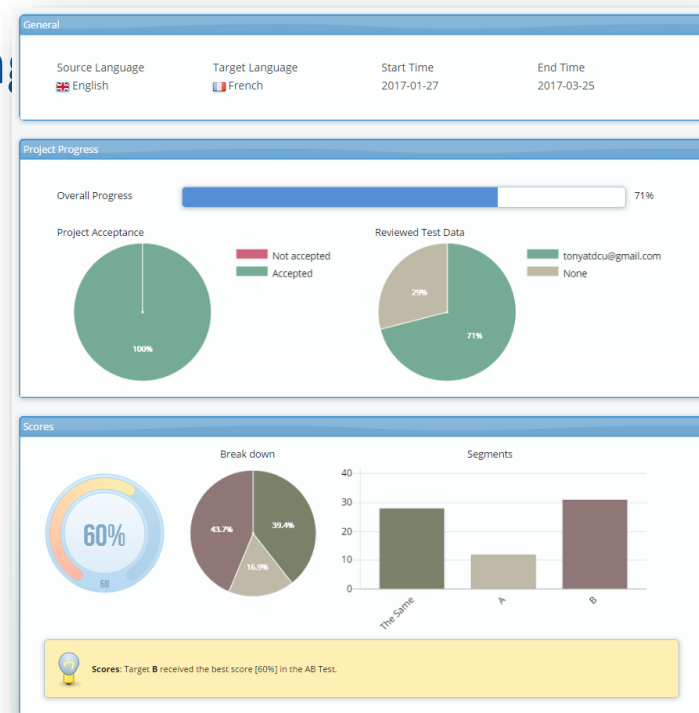
This year, over 4,000 Canadian veterans commemorate the fifty-fifth anniversary of VE Day by returning to Holland, a place that holds a special meaning for them.

Cette année, plus de 4 000 anciens combattants canadiens commémorent le cinquante-cinquième anniversaire du jour de la victoire en Europe en retournant en Hollande, un endroit qui a une signification particulière pour eux. Better

Cette année, plus de 4 000 anciens combattants canadiens commémorent le cinquante-cinquième anniversaire du jour de la victoire en Europe en retournant en Hollande, un endroit qui a une signification particulière pour eux. Better

The Same

Save & Finish Cancel



Empirical evaluation

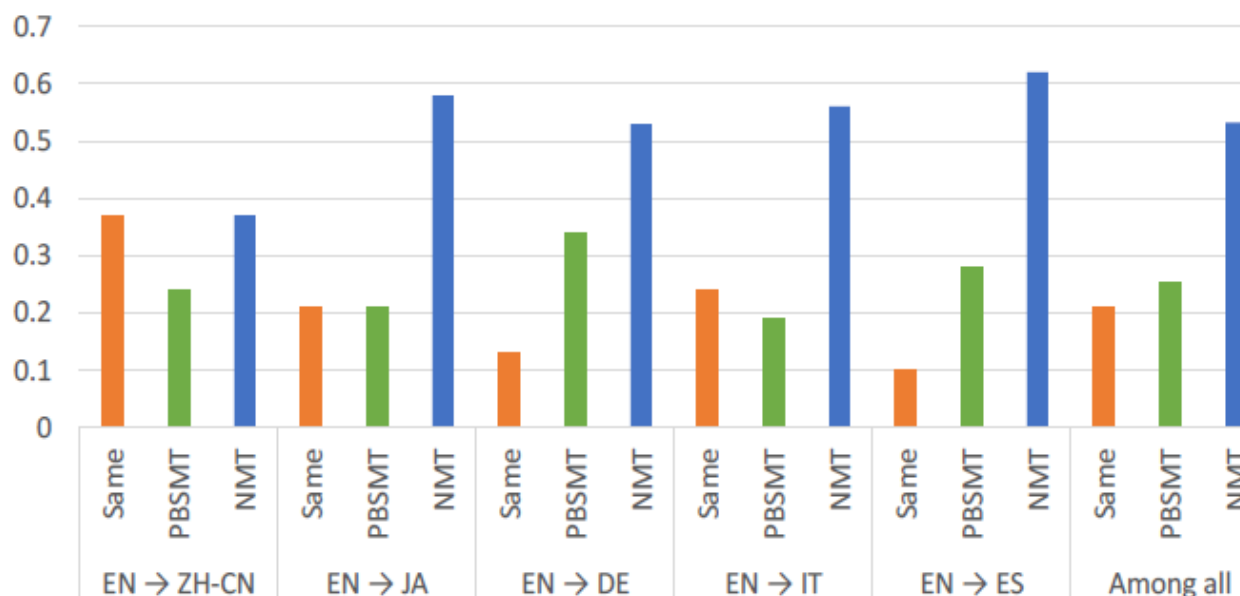
- **Human evaluation:**

- Side-by-side (with KantanLQR / ABTesting)
- 200 sentence triples
- Native speakers of the target language; proficient in English

Empirical evaluation

- **Human evaluation:**

- Side-by-side (with KantanLQR / ABTesting)
- 200 sentence triples
- Native speakers of the target language; proficient in English



Empirical evaluation

- **BLEU analysis on the AB Test results**

- Set of triplets for which the translation produced by the NMT engine was considered better.
- From this set count the translations that are scored by BLEU **lower** than their PBSMT counterparts.
- Do the same for the PBSMT translations.

	EN-ZH-CN	EN-JP	EN-DE	EN-IT	EN-ES	Average
NMT	40%	59%	55%	34%	53%	48%
PBSMT	12%	0%	9%	9%	0%	6%

Future work

- **Perform further evaluation:**
 - Error analysis
 - Other language pairs
- **Optimise the training pipeline**
- **Improve quality evaluation**

- **Acknowledgements:**

Xiyi Fan, Ruopu Wang, Wan Nie, Ayumi Tanaka, Maki Iwamoto, Risako Hayakawa, Silvia Doehner, Daniela Naumann, Moritz Philipp, Annabella Ferola, Anna Ricciardelli, Paola Gentile, Celia Ruiz Arca, Clara Beltr.

The University College London, Dublin City University, KU Leuven, University of Strasbourg, and University of Stuttgart.



KantanMT.com

No Hardware. No Software. No Hassle MT.

Dimitar Shterionov: dimitars@kantanmt.com

Pat Nagle: patn@kantanmt.com

Laura Casanellas: laurac@kantanmt.com

Riccardo Superbo: riccardos@kantanmt.com

Tony O'Dowd: todyod@kantanmy.com

KantanLabs: labs@kantanmt.com

General: info@kantanmt.com

Thank you...