



Preserving MT Quality for Content With Inline Tags

Grigory Sapunov,
* Konstantin Savenkov,
Pavel Stepachev

AGENDA

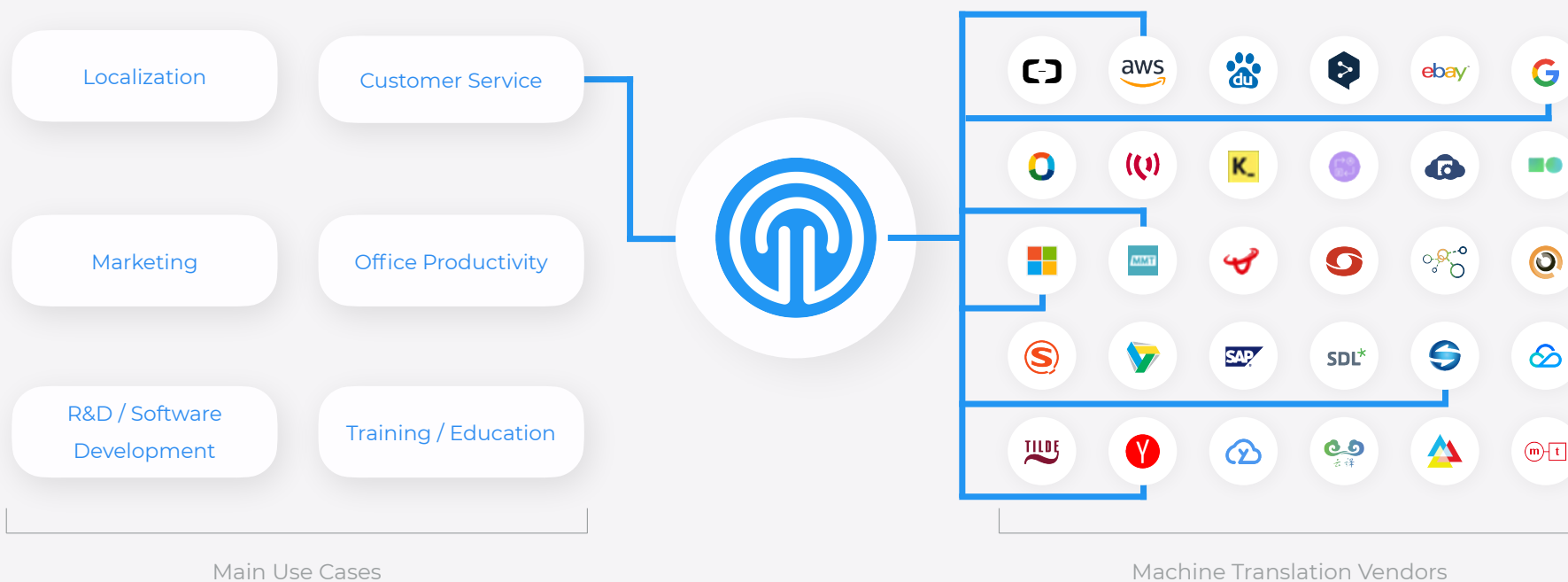
Why tags and placeholders are important?

-
- MT + tags = it's complicated
-
- Intento Solution: **Smart Tag Handling**
-
- Experimental setting
-
- Experimental results
-
- Current status and next steps

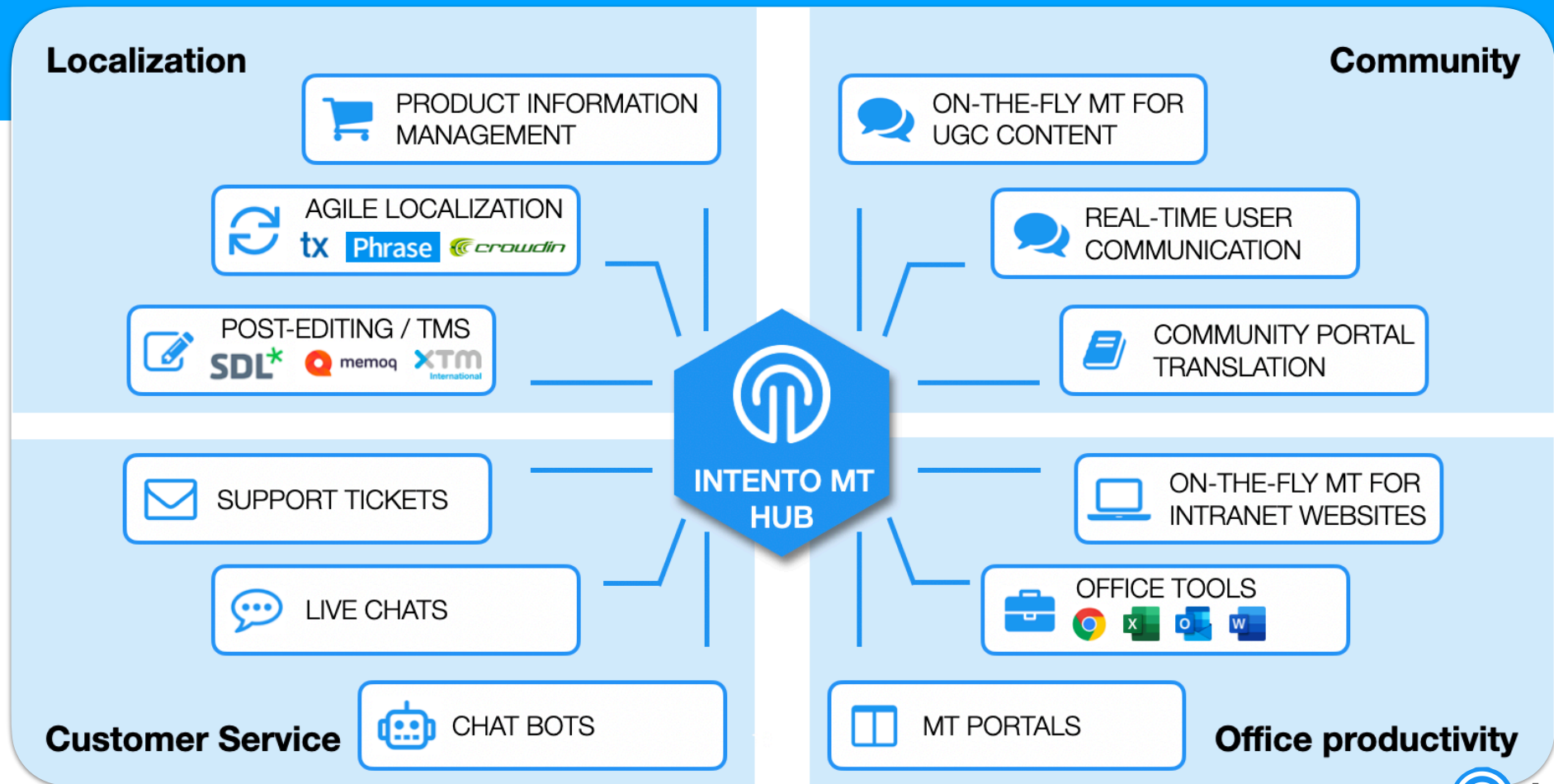


ABOUT INTENTO

Intento MT Hub integrates AI/ML models from many vendors into the business processes, choosing the best-fit combination for every use case



MULTI-PURPOSE MT

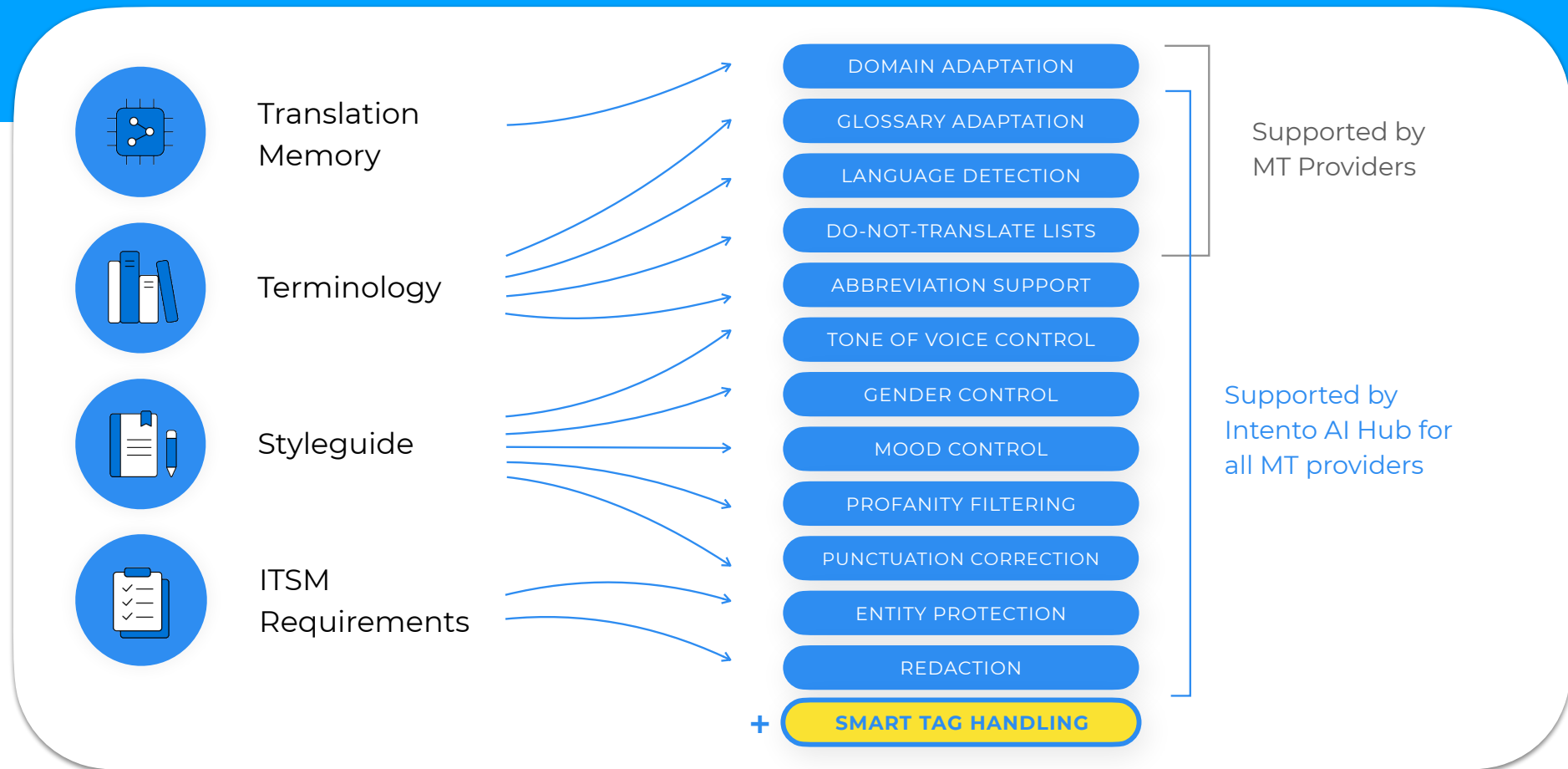


MT REQUIREMENTS MATRIX

EVERY CASE HAS ITS OWN NEEDS

	large text translation	batch translation	latency and jitter	tolerance to bad source	language detection	tag support	multilingual source	profanity control	metadata protection	entity protection	custom terminology	tone of voice control
Post-editing / TMS		●				●				●		
Support tickets	●				●		●	●	●	●	●	
Live chats			●	●	●			●		●	●	●
Subtitle translation			●	●	●	●		●		●	●	●
On-the-fly UGC		●	●	●	●	●	●	●		●		●
Real-time communication			●	●					●			●
Knowledge bases	●					●		●		●		

USE-CASE SPECIFIC MT FEATURES



SMART TAG HANDLING

Even Custom NMT does not always deliver

I had a delivery recently in
Orlando



Ich hatte vor kurzem eine Geburt
in Orlando

NMT + TAGS

IT'S COMPLICATED

Inconsistent across MT providers and language pairs.

—

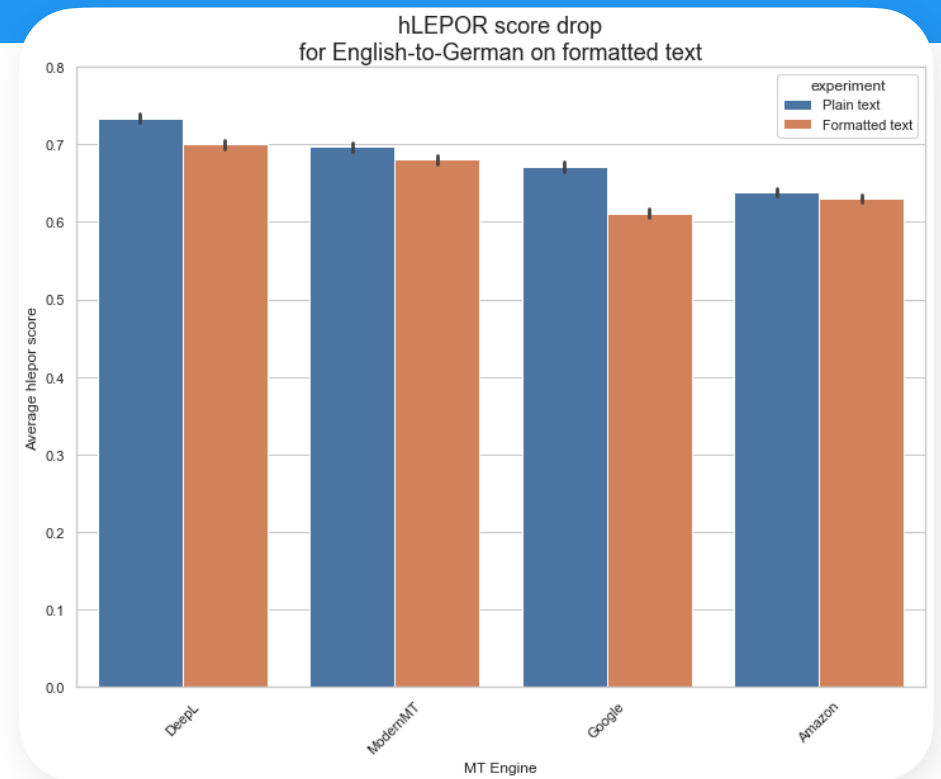
Customized models may fall back onto baseline because of tags.

—

Placeholders are impossible for MT to interpret.

—

Glossaries also break as they often rely on tags.



CURRENT SOLUTIONS

Raw MT: 🙄

MTPE: either spend post-editor time on editing broken language, or remove tags and spend post-editor time on putting them back.

Our primary use-case: **video translation** (mistreated tags are critical, editing them is complicated)

MOVING TAGS OUT OF THE EQUATION

```
I had a delivery recently  
<timestamp class='timestamp'  
  start='00:00:13,230'  
  end='00:00:17,690' />  
in <ph/>
```

MOVING TAGS OUT OF THE EQUATION

(1) Removing inline tags

I had a delivery recently in <ph/>



```
<timestamp  
class='timestamp'  
start='00:00:13,230'  
end='00:00:17,690' />
```


MOVING TAGS OUT OF THE EQUATION

(2) Filling placeholders with generative models

I had a delivery recently in **New York**

```
<timestamp  
class='timestamp'  
start='00:00:13,230'  
end='00:00:17,690' />
```

<ph/>

MOVING TAGS OUT OF THE EQUATION

(3) Translating plain text

I had a delivery recently in **New York**



Ich hatte kürzlich eine Lieferung in New York

```
<timestamp  
class='timestamp'  
start='00:00:13,230'  
end='00:00:17,690' />
```

```
<ph />
```

MOVING TAGS OUT OF THE EQUATION

(4) Performing word alignment

I had a delivery recently in **New York**

Ich hatte kürzlich eine Lieferung in New York

```
<timestamp  
class='timestamp'  
start='00:00:13,230'  
end='00:00:17,690'/>
```

```
<ph/>
```

MOVING TAGS OUT OF THE EQUATION

(5) Putting tags back

I had a delivery recently in **New York**

Ich hatte kürzlich eine Lieferung<timestamp
class='timestamp'
start='00:00:13,230'
end='00:00:17,690' />
in <ph/>

EXPERIMENTS

EXPERIMENTS

TWO EXPERIMENTS

A: HTML FORMATTING

How much MT quality suffers from simple HTML tags?

Using Smart Tag Handling to put tags back after MT

B: PLACEHOLDERS

How much MT quality suffers from words replaced by placeholders?

Does translating text w/o placeholders help?

Using Smart Tag Handling to put placeholders back after MT

Does expanding placeholders help?

EXPERIMENTS

ORIGINAL DATASET

EN-DE corpus from TAUS

—

Domain - Financial Services

—

1955 segments

—

> 5 tokens per segment

The investigation confirmed the complainant's legal claim that the C-57 Amendment to the Canadian Trade-Marks Act violates Articles 23.1 and 2 as well as Article 24.3 (the so-called standstill clause) of TRIPS and that such infringements cannot be justified on the basis of the exception under Article 24.6 of TRIPS.

Die Untersuchung bestätigte die rechtliche Behauptung des Antragstellers, das Gesetz C-57 zur Änderung des kanadischen Handelsmarkengesetzes verstoße gegen Artikel 23 Absätze 1 und 2 sowie Artikel 24 Absatz 3 (die so genannte Stillhalteklause) des TRIPS, und dieser Verstoß könne nicht durch die Ausnahmeregelung des Artikels 24 Absatz 6 des TRIPS gerechtfertigt werden.


A - TAGGING

DATA PREPARATION A - TAGGING

1-3 tag entries per segment

—
tags: 1-place (img, br) or 2-
place (span, i, em, a, b,
strong, u, s)

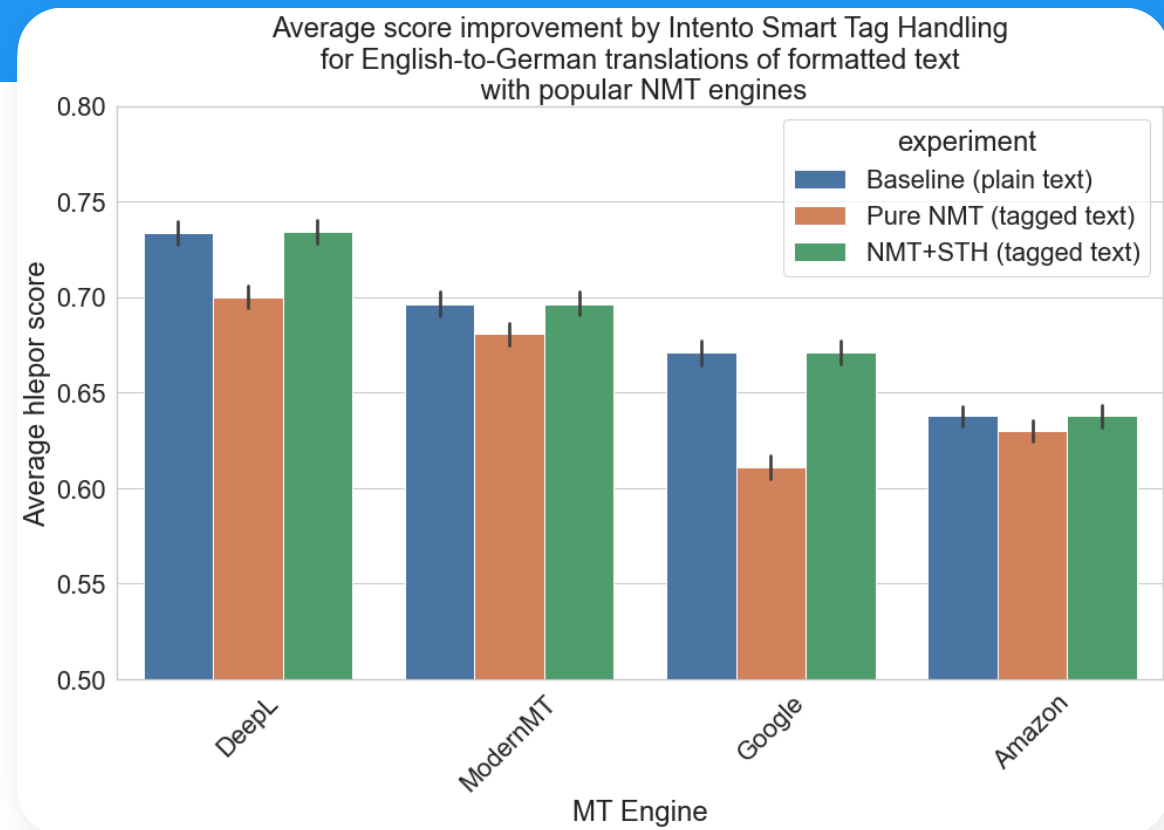
—
nesting: 1-3 levels

The investigation confirmed the complainant's legal **** claim that the C-57 Amendment to the Canadian Trade-Marks Act violates Articles 23.1 and [](https://example.com/index.html) 2 as well as ~~Article 24.3~~ (the so-called standstill clause) of TRIPS [](#) and that such infringements cannot be justified on the basis of the exception **** under Article 24.6 of  [src="https://example.com/image.png"](https://example.com/image.png) alt="Some image"/> TRIPS.

A - TAGGING SCORING

Calculate hLEPOR score for:

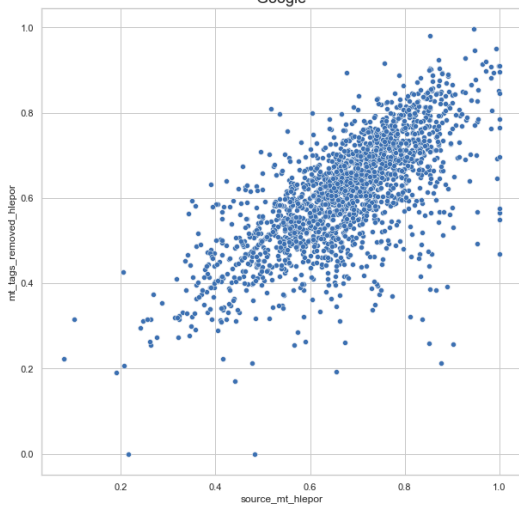
-
- (1) Plain text NMT
- (2) Tagged text NMT after tag removal
- (3) Tagged text NMT+STH after tag removal



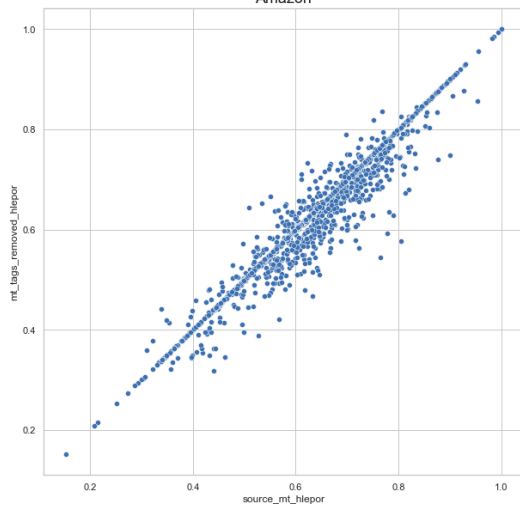
A - TAGGING

SEGMENT DEGRADATION DUE TO TAGS

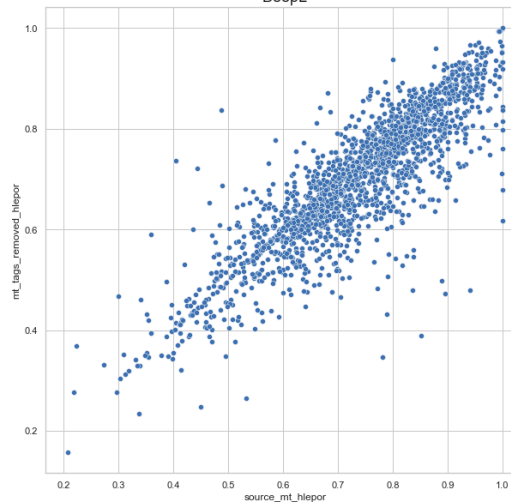
Score change after adding inline tags
Google



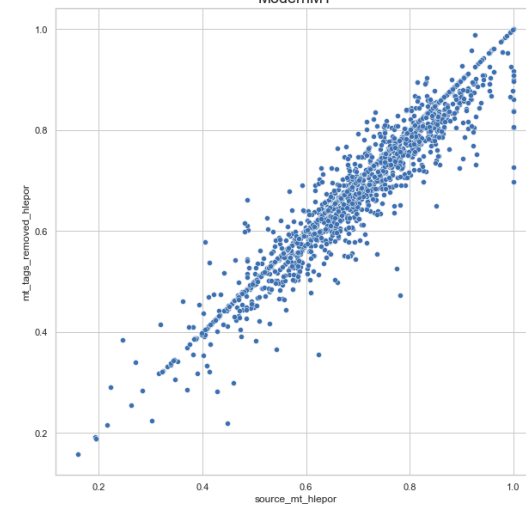
Score change after adding inline tags
Amazon



Score change after adding inline tags
DeepL



Score change after adding inline tags
ModernMT

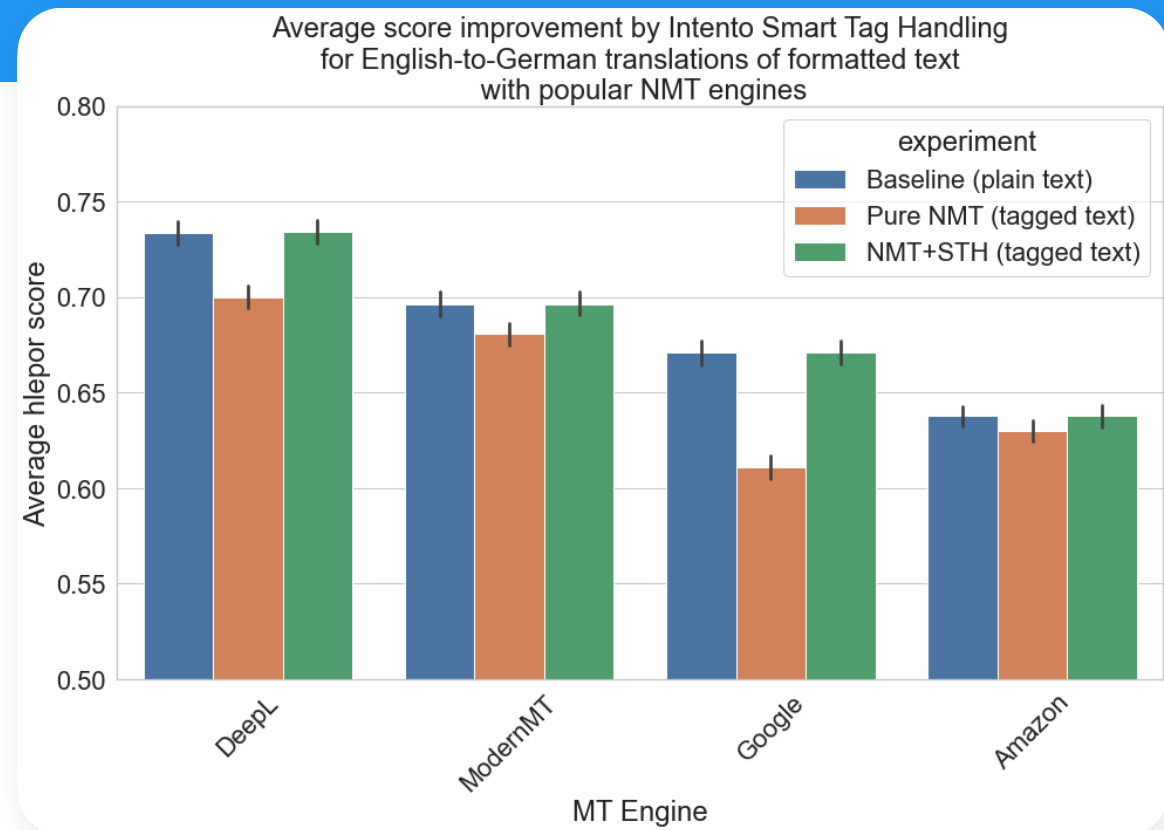


A - TAGGING DISCUSSION

Even innocent HTML tags degrade NMT quality (as of today).

The way to improve the quality is to translate text with tags removed and insert them back after MT

Benefits: (1) same level of translation quality as plain text, (2) post-editor does not spend time to move tags, (3) natively integrated into the existing AVT workflow.



B - PLACEHOLDERS

DATA PREPARATION

Same dataset, 367 segments with DNT.

—
Non-translatables replaced with placeholder tags.

The investigation confirmed the complainant's legal claim that the `<ph/>` Amendment to the Canadian Trade-Marks Act violates Articles 23.1 and `<ph/>` as well as Article 24.3 (the so-called standstill clause) of `<ph/>` and that such infringements cannot be justified on the basis of the exception under Article 24.6 of `<ph/>`.

Die Untersuchung bestätigte die rechtliche Behauptung des Antragstellers, das Gesetz `<ph/>` zur Änderung des kanadischen Handelsmarkengesetzes verstoße gegen Artikel 23 Absätze 1 und `<ph/>` sowie Artikel 24 Absatz 3 (die so genannte Stillhalteklause) des `<ph/>`, und dieser Verstoß könne nicht durch die Ausnahmeregelung des Artikels 24 Absatz 6 des `<ph/>` gerechtfertigt werden.

B - PLACEHOLDERS

DATA PREPARATION

Same dataset, 367 segments with DNT.

—
Non-translatables replaced with placeholder tags.

—
Placeholder tags are expanded with dummy values using multilingual generative language model.

The investigation confirmed the complainant's legal claim that the `<ph/>` Amendment to the Canadian Trade-Marks Act violates Articles 23.1 and `<ph/>` as well as Article 24.3 (the so-called standstill clause) of `<ph/>` and that such infringements cannot be justified on the basis of the exception under Article 24.6 of `<ph/>`.

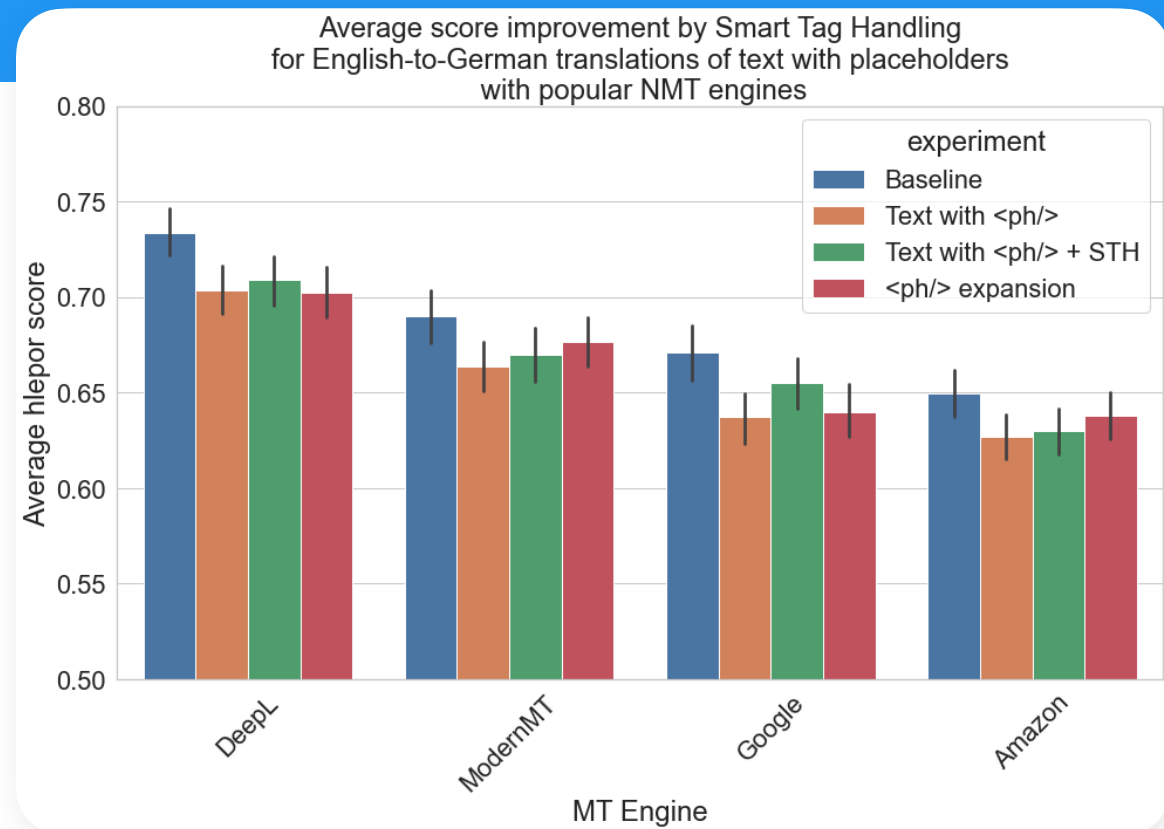
The investigation confirmed the complainant's legal claim that the `Second` Amendment to the Canadian Trade-Marks Act violates Articles 23.1 and `23.2` as well as Article 24.3 (the so-called standstill clause) of `NAFTA` and that such infringements cannot be justified on the basis of the exception under Article 24.6 of `NAFTA`.

B - PLACEHOLDERS

SCORING

Calculate hLEPOR score for:

- (1) Plain text NMT with removed DNT vs reference translation with removed DNT (Baseline)
- (2) Text with <ph/> vs reference (Raw NMT)
- (3) Text with <ph/> vs reference (NMT+STH)
- (4) Text with expanded <ph/> vs reference

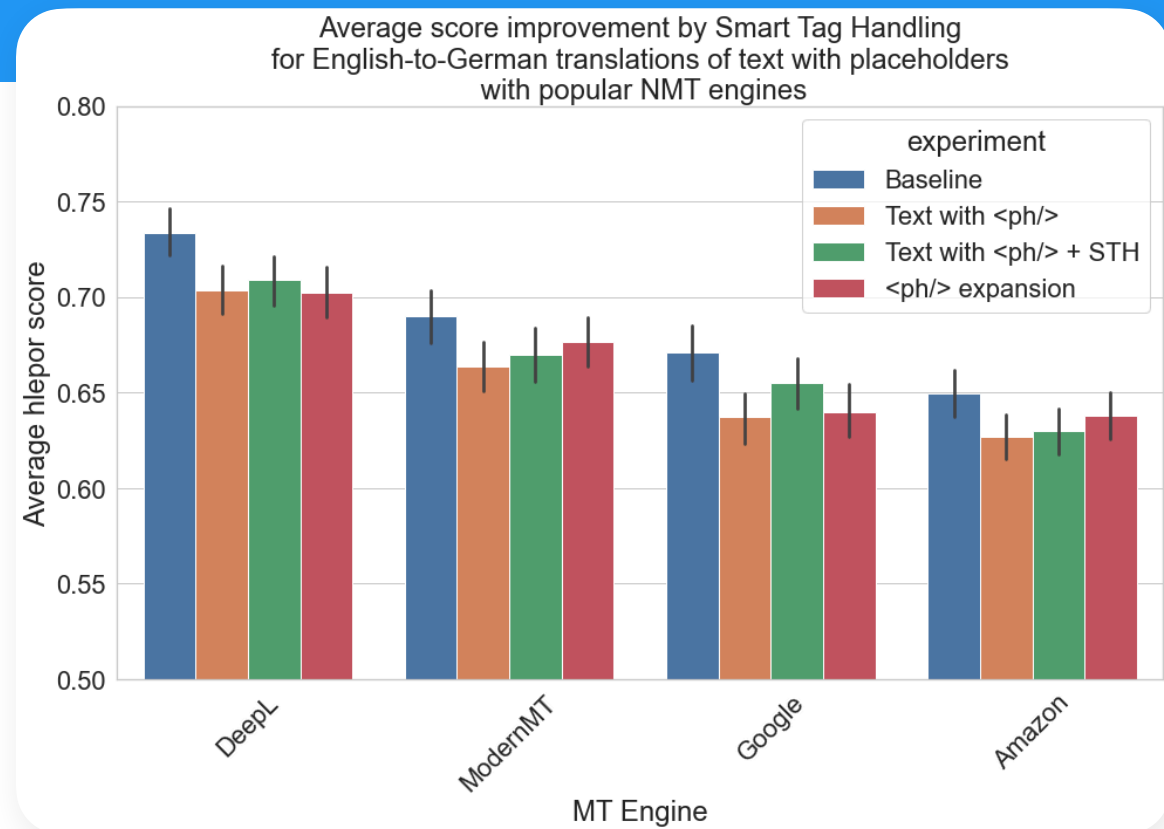


B - PLACEHOLDERS DISCUSSION

Adding placeholders significantly decreases MT quality for all MT engines.

Using STH for <ph/> improves MT quality. Level of improvement depends on how well MT engine deals with incomplete sentences vs. sentences with <ph/> tags.

Expanding placeholders further helps for some engines (ModernMT and Amazon), should not be used for others (Google and DeepL).



CONCLUSIONS

Even innocent HTML tags degrade NMT quality (as of today).
Placeholders too.

—

The way to improve the quality is to translate text with tags removed
and insert them back after MT.

—

Also, this is a must for MT engines that are best for certain
languages, but lack tag support (Tencent, Baidu, Naver, etc)

—

For placeholders, removing placeholders from translation altogether
also improves the MT quality.

—

Placeholder expansion helps for some MT engines, for others it needs
improvement.

CURRENT STATUS

Available as an automated post-editing via API for selected customers.

—

The main use-case so far - subtitle translation in TMS, to reduce time spent on both text editing and timestamp re-placement.

—

We are planning to evaluate ROI (cost and TAT decrease) for AVT with one of our customers, we'll keep you posted :-)

REMAINING ISSUES AND NEXT STEPS

Our tag placement algorithm works decently for single-position tags (timestamps, img, br).

—

Putting back deeply nested HTML structures requires further improvement.

—

Placeholder expansion requires improvement to avoid using tags to track the position of the expanded `<ph/>`.

Using MT for inline tags?
Let us know!
ks@inten.to