
Multi-Domain Adaptation in Neural Machine Translation Through Multidimensional Tagging
E. Stergiadis, S. Kumar, F. Kovalev, P. Levin
Summer 2021
Booking.com


Motivation
Booking.com

# NMT in the wild.

Most industrial NMT systems to seek to serve a niche market. E.g think of what some practitioners might be interested in translating:

- Booking.com
- An E-Commerce company
- A health infrastructure provider
- ...

Booking.com

# How do people serve a niche? Fine-tuning!

We trade a **generalist** model (reasonably good in translating everything) for a **specialised** one (really good at translating our niche, likely worse at everything else).

Booking.com

# But what if we want to serve several niches?

1. Property Descriptions

*"Featuring an outdoor pool, **Kastro Hotel** is located in **Agios Kirykos Village**. It offers air-conditioned rooms and free WiFi throughout. **Agios Kirykos Port** is just **200 m** away."*

- Factual
- Formal
- 3rd person
- **Named Entities**

Booking.com

# But what if we want to serve several niches?

2. Guest Reviews

*"to enter bathroom you have to open inside the door step into corner (preliminary removing small plastic chair standing there ) between washstand and toilet than make a pirouette around and try to close the door from inside)))) a little bit tricky((("*

- Emotional
- Informal
- Often 1st person
- Grammatical Errors + Slang → Much higher OOV %

Booking.com

# But what if we want to serve several niches?

2. Messages (fake example)

" - *Hi, my name is Manos. Can I bring my dog pls is good dog promise*

- *Dear Manos, thank you for reaching…. According to the booked property's policy… we hope we have informed you sufficiently, with warm regards….*"

- PII data
- Varies from very informal (guest) to very formal (CS agent)
- Often full of typos

Booking.com

# But what if we want to serve several niches?

## We could...

**Build one model for each domain**

- Expensive to make and maintain (if we serve 40 languages we need 240 models)
- Booking expands into insurance, flights and experiences offering. Do we build 240 new models?

**Build a single model for all**

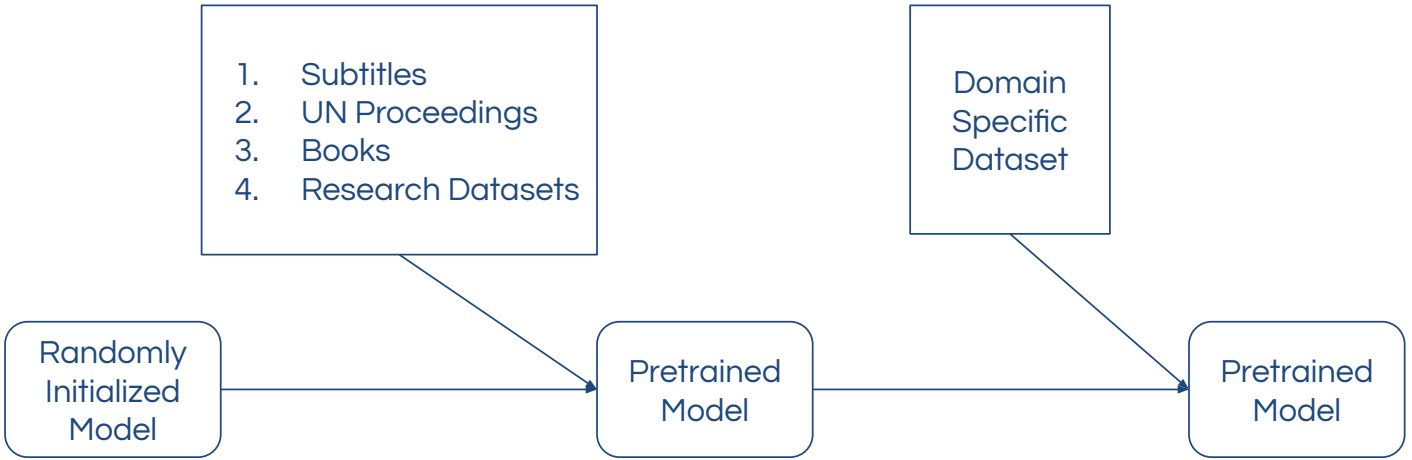
- Much more scalable but naively concatenating datasets yields a model worse than specialised ones

Can we get the best of both worlds? Enter **MDT!**

Booking.com


Background
Booking.com

# Fine-tuning




Booking.com

# NMT in the wild

| Generic Dataset | In Domain Dataset |
| --- | --- |
| Cheap | Expensive |
| Big | Small |
| Irrelevant | Relevant |

Booking.com

# NMT in the wild

| Generic Dataset | In Domain (Parallel) Dataset | In Domain (Monolingual) Dataset |
|---|---|---|
| Cheap | Expensive | Cheap |
| Big | Small | Big |
| Irrelevant | Relevant | Relevant |

Booking.com

# Back-Translation

- Popularised in 2015 (R. Sennrich et al)
- Improved by adding Noise (S. Edunov et al)
    - … **or by adding a tag**! (I. Caswell et al)

Booking.com

# Multi Dimensional Tagging

# MDT basic idea

Tagged back-translation uses a tag to pass **source distribution level metadata** when translating a sentence: whether or not the source text has been generated by a human or back-translation.

Other haved used tags to pass different kinds of meta-information:

- Language (M. Johnson et al)
- Domain (C. Kobus et al)
- Style (R. Sennrich)

**But what if we pass many?**

Booking.com

# How does it look in practice?

*This is a **real** review written by a portuguese person*

"<PT> **<REAL>** <REVIEW> A localização é boa, com acesso fácil aos principais locais de interesse."

*This is a **real** property description written by a brazilian person*

"<BR> **<REAL>** <PROPERTY> TV de tela plana e conta com estacionamento privativo gratuito no local."

*Observe that the information conveyed by each tag is orthogonal to that of others (no redundancy)*

Booking.com

# A note on applicability

- Our method is model-agnostic as it pertains to data preprocessing rather than architecture specific tweaks.
- Some form of global interaction attention might be required for the method to work as only then can each token independently query the tag.
- Global interactions exist explicitly in Transformers that are SOTA today but also implicitly in RNNs and sufficiently deep CNNs.

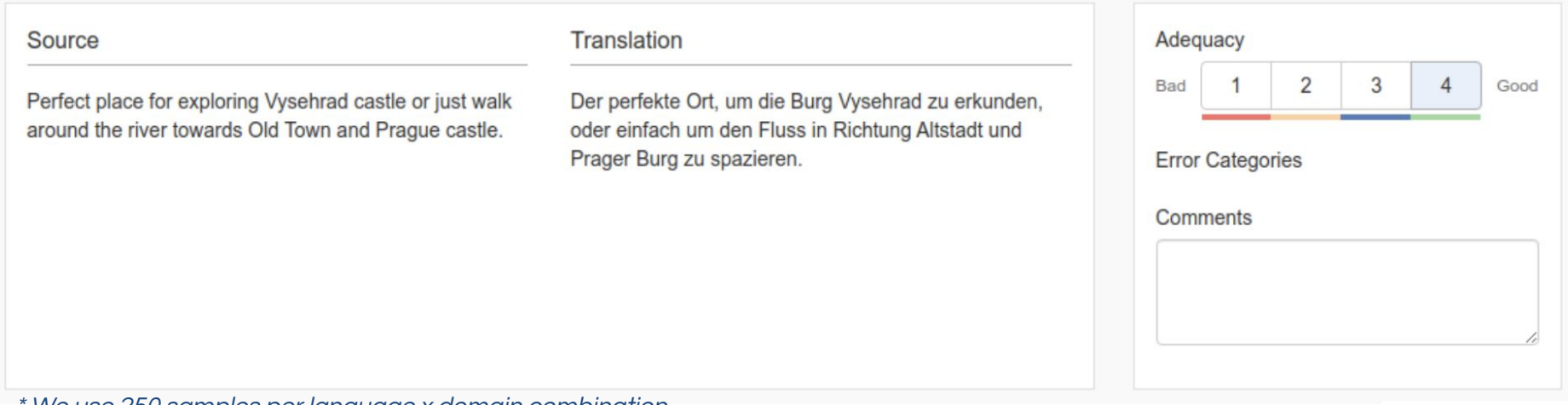
Booking.com

# Experiments & Results

# Evaluation

BLEU is the industry standard metric when evaluating MT systems. However multiple studies have shown its weaknesses even against other automatic metrics that better correlate with human scores. Most recently for example: T. Kocmi et al

To address this, we obtain and report human evaluation scores for different systems*.




** We use 250 samples per language x domain combination.*

Booking.com

# Baselines

We experiment with 3 languages, each using a different alphabet. For each, we report scores from:

1) The base model
2) A domain specific fine-tuned model* (which is our main baseline)
3) The MDT model (common for all domains).

Both (2) and (3) fine-tune starting from the same checkpoint of (1).

*In the past we had extensively experimented with single-domain fine-tuning strategies. We report scores of the one that beat all other alternatives and corresponds to the method called "top10" S. Edunov et al. Namely, during back-translation each token is decoded with the softmax strategy where only the top 10 candidates are considered. In fact this baseline had been our live model for more than 1 year across 16 languages!

Booking.com

# Does it work?

| | Reviews | | | Messaging | | | Descriptions | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AR | DE | RU | AR | DE | RU | AR | DE | RU | AR | DE | RU |
| **Human score** | | | | | | | | | | | | |
| Base model | 3.65 | 3.73 | 3.50 | 3.27 | 3.44 | 3.18 | 2.67 | 3.28 | 2.95 | 3.20 | 3.48 | 3.21 |
| +top10 | **3.75 (+.10)** | 3.80 (+.07) | 3.57 (+.07) | 3.36 (+.09) | 3.65 (+.19) | 3.53 (+.35) | 3.02 (+.35) | 3.70 (+.42) | 2.95 (+.00) | 3.38 (+.18) | 3.71 (+.23) | 3.47 (+.14) |
| +MDT | 3.72 (+.07) | **3.88 (+.15)** | **3.62 (+.12)** | **3.49 (+.22)** | **3.78 (+.34)** | **3.53 (+.35)** | **3.20 (+.53)** | **3.73 (+.45)** | **3.04 (+.09)** | **3.47 (+.27)** | **3.80 (+.31)** | **3.40 (+.19)** |
| **BLEU score** | | | | | | | | | | | | |
| Base model | 42.95 | 43.63 | 38.25 | 39.01 | 44.18 | 41.18 | 45.00 | 45.97 | 38.92 | 42.32 | 44.60 | 39.45 |
| +top10 | **42.95 (+0.00)** | 44.99 (+1.36) | 38.35 (+0.10) | 41.93 (+2.92) | **50.19 (+6.01)** | 41.15 (-0.03) | 45.35 (+0.35) | **50.98 (+5.01)** | 37.84 (-1.08) | 43.41 (+1.09) | 48.72 (+4.13) | 39.11 (-0.34) |
| +MDT | 42.61 -0.34 | **46.34 (+2.71)** | **41.12 (+2.87)** | **47.09 (+8.08)** | 49.85 (+5.67) | **43.19 (+2.01)** | **46.54 (+1.54)** | 50.84 (+4.87) | **39.14 (+0.22)** | **45.41 (+3.09)** | **49.01 (+4.41)** | **41.15 (+1.70)** |

Booking.com

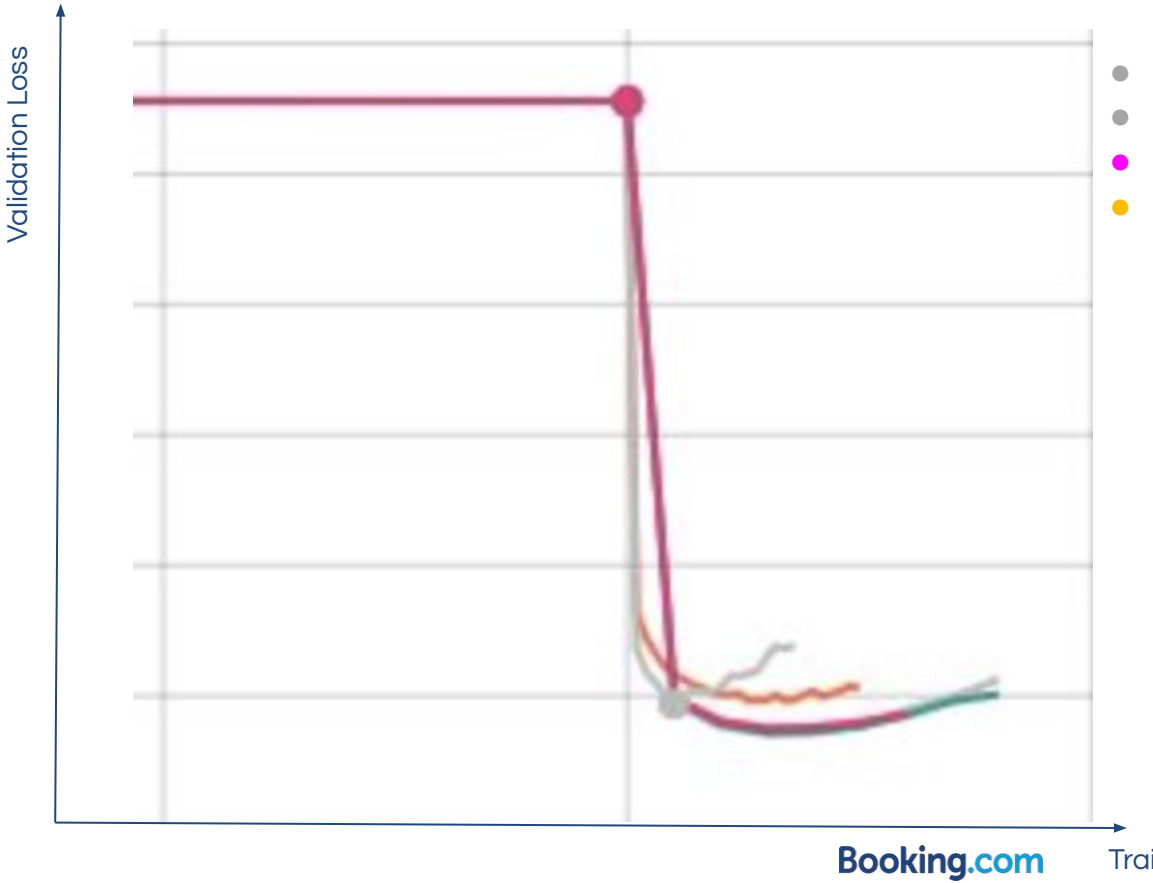
# Why tags? Just put all domains together

| | Human score | BLEU score |
|---|---|---|
| **Reviews** | | |
| MDT Model | 3.88 | 46.34 |
| (-tags) | 3.82 (-.06) | 44.24 (-2.10) |
| **Messaging** | | |
| MDT Model | 3.78 | 49.85 |
| (-tags) | 3.48 (-.30) | 49.21 (-0.64) |
| **Descriptions** | | |
| MDT Model | 3.73 | 50.84 |
| (-tags) | 3.80 (+.07) | 49.79 (-1.05) |
| **Average** | -.10 | -1.26 |

*Above we average scores across languages (and in the last row across domains too)*

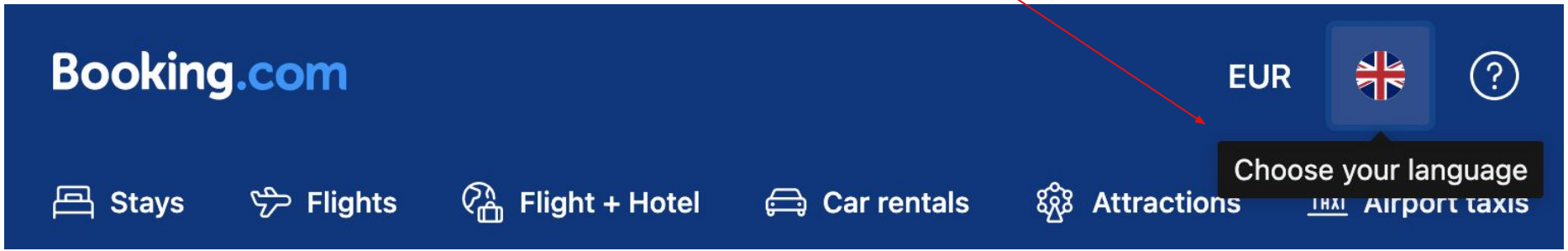
Booking.com

# Why tags? Just put all domains together



# Conclusion

MDT works at least as well as domain specific models while addressing 3 domains, therefore reducing the number of production models by the same factor. It now powers all those domains across 16 languages in Booking.com. **Try it out!**



# Future Work

1. We empirically evaluate MDT using a 2D tags (origin + domain). It would be interesting to detect the method's limits: how many tags and at what level of granularity can we insert before results deteriorate?

2. How does the method fair when new domains (unseen during training) are introduced by the business? Is the widened training distribution enough to render the model robust to new domains, or is a second round of fine-tuning required?

Booking.com