Booking.com NMT a OpenNMT use case

Talaat Khalil, Data Scientist

OpenNMT Workshop Paris, March 2018

Presentation Outline

- MT use cases at Booking.com
- Pipeline and Standard Architecture
- Evaluation pipeline
- Domain Adaptation for User Generated Content (UGC)
- Translation challenges
- Production pipeline
- Recommendations & feature requests
- References

Use cases at Booking.com

2/3 of daily bookings on Booking.com is made in a language other than English

… thus it is important to have **locally relevant content at scale**

How Locally Relevant?

Allow partners and guests to **consume and produce content in their own language**

- Hotel Descriptions
- Customer Reviews
- Customer Service Support

Why At Scale?

- **One Million+ properties** and growing very fast
- **Frequent change requests** to update the content
- **43 languages** and more
- New user-generated **customer reviews / tickets** every second

Main use cases at Booking.com

- Hotel & Room Descriptions
  - 50% human translation coverage
  - 90% demand coverage
  - Average of 10M parallel sentences for high demand languages

- User Reviews
  - No Translation coverage
  - No In-Domain data

Pipeline and Standard Architecture

Preprocessing	Training	Translation	Post Processing
- Data Cleaning - Handling numbers and Named Entities* - OpenNMT preprocessing + extra domain related features**	-Seq to seq arch, described in the following slide -Domain Fine-Tuning*	- Determining best combination to translate - OpenNMT Translate	- OpenNMT post processing - Numbers and Named Entities post processing* - Predicting errors in MT text

Evaluation
- Automatic Evaluation - Human Evaluation

* For some use cases
** Depending on language / use case

Pipeline and Standard Architecture

NMT

Data Preprocessing	
Input text unit	Lowercased BPE
Tokenization	Aggressive, with case features
Max. sent length	50 units
Vocab Size	30,000-50,000 * * Joint or Separate

Architecture ** Variant of (Bahdanau et al)	
Input dim	1000
RNN dim	1000
# of hidden layers	Encoder: 4 Decoder: 4
Attention mechanism	Global
RNN Type	LSTM ** Bidirectional encoder
Residual connections	Yes

Optimization ** Standard pipeline	
Optimizer	SGD
Learning rate decay	0.7
Decay strategy	Validation perplexity increase or Epoch > 20
Stopping Criteria	Based on validation perplexity
Dropout rate	0.3
Max Batch size	120

Others	
Inference Beam size	5
GPU	Nvidia P100

Evaluation Pipeline

- BLUE score Evaluations for model development

- Human Evaluation:
  - Adequacy and Fluency checks before model deployment
  - Periodic checks or random samples
  - Biased sample evaluation:
    - Score recent production translations using our in-house error detection model
    - Send lowest scored sentences for human evaluation

Domain Adaptation for UGC (Data)

- Little to no in-domain data.

- In addition to our hotel descriptions data, available external open data is used including data from:
  - Movie subtitles
  - Wikipedia
  - TED talks
  - New commentary
  - EuroParl

Domain Adaptation for UGC (Data)

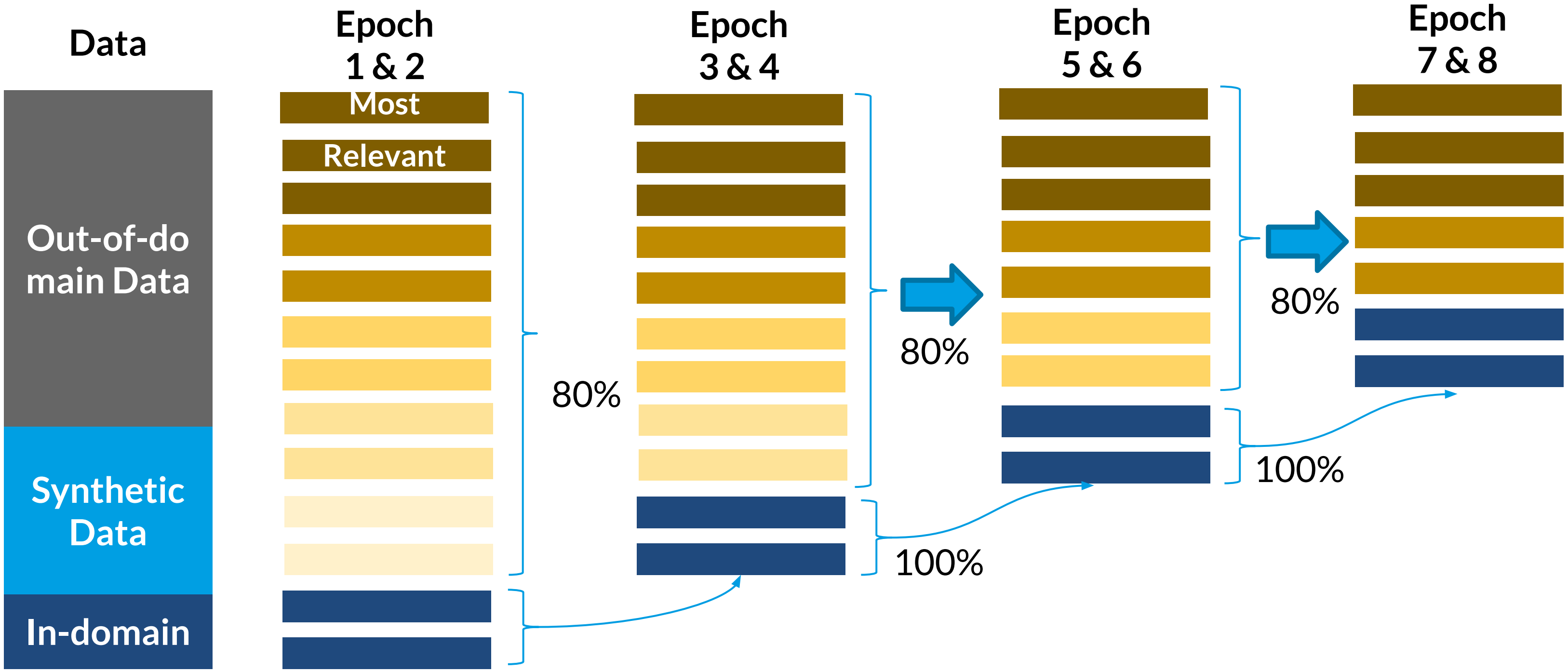
- Synthetic Data

- We then use either Gradual Downsampling (Wees et al.) or Fine Tuning for domain adaptation

Gradual Downsampling approach

Data	Idea	Methodology
Out-of-Domain Data	Use in-domain and general data language models to select most relevant sentences from general data	Bilingual Cross Entropy Difference (Axelrod et al) - To select sentences that are most similar to in-domain but different to out-of-domain.
Synthetic Data	Use large amount of mono-lingual in-domain data to create some synthetic in-domain data	Back translate target language in-domain data into source by reversing our MT model (Sennrich et al.). (Ranked Randomly)
In-domain Data	Create a small amount of in-domain corpus as well, to test for additional impact	Human Translation

Gradual Downsampling approach



Fine-Tuning Approach

- Train with Open data + in house out of domain data.
- Wait until the model converges
- Fine tune with few in-domain data samples (reviews)

Domain Adaptation for UGC

Gradual Downsampling	Fine-Tuning
Faster iterations	Takes time to get the General Model trained
Trained for specific use case from the beginning	Can be adapted to multiple use cases
Applicable without In-domain parallel data	Needs In-domain parallel data
Less accurate	More Accurate

- General purpose model or Gradual downsampling model? (without in-domain data)
  - Not Answered yet!

Translation Challenges

- Numbers Translations
  - Solution: Placeholders for distances, time, currency, date, etc,.
- Named Entities (NE) Translations
  - Solution: NE placeholders and NE tagging models (still tricky)
- Rare words Translations
  - On average our system is better than a general purpose
  - Not always the case when it comes to rare words, even if you have millions of in-domain data
  - Training with general purpose data and fine tuning with domain data helps

Translation Challenges

- Repetitions and Omissions:
  - Incorporation of more general language models for encoder/decoder initialization and training (Ramachandran et al.)
  - OpenNMT support is needed!

- Inaccurate source language (grammar and punctuation)

Translation Challenges

- Context handling
  - Mainly gender issues:
    - Property type co-reference can be male/female in some languages. This depends mainly on the property type gender in the previous sentence
    - We use property type features for some languages
  - Exploiting context from previous/next sentences (Bawden et al.)
  - OpenNMT Support is needed!

Production Pipeline

- ZeroMQ server provides more flexibility however it's not compatible with our HTTP based infrastructure.

- We use our batched version of the REST server:
  - We have our own production preprocessing functionalities.
  - The OpenNMT REST server is used then for the translation.
  - It was necessary to have this separation to be able to add our crafted features during preprocessing and use the server only for translation.

Recommendations & Feature requests

- Instance weighting
  - Explicit Instance weighting that is incorporated in the loss function with the possibility of having negative weights (wrong translations)
  - Resample data between epochs/batches and adapt instance weights based on validation performance
  - Wang et al, showed some instance weighting and adaptive weighting approaches that could be of inspiration

Recommendations & Feature requests

- Always start with models trained on all the available data

- Our recent experiments shows significantly better results when we used the “Big Transformer” architecture as described by Vaswani et al. (lua version support?)

Recommendations & Feature requests

- Multi-encoder support
  - Make use of our available features for sequence to sequence problems (context, images, etc,.)

- More support for Tensorflow version is more crucial for production systems
  - Tensorflow has more proper versioning unlike torch
  - Installation is more easier for production (lua torch is more optimized for research and easier to install for user level)
  - Tensorflow is more supported in more production environments

References

- Bahdanau et al. Neural Machine Translation by Jointly Learning to Align and Translate, ICLR 2015.
- Wees et al. Dynamic Data Selection for Neural Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1411–1421 Copenhagen, Denmark, September 7–11, 2017
- Sennrich et al. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 86–96, Berlin, Germany, August 7-12, 2016.
- Axelrod et al. Domain adaptation via pseudo in-domain data selection. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 355–362, July 2011.
- Vaswani et al. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- Wang et al. Instance Weighting for Neural Machine Translation Domain Adaptation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1482–1488 Copenhagen, Denmark, September 7–11, 2017.

References

- Bawden et al. Evaluating Discourse Phenomena in Neural Machine Translation. In Proceedings of NAACL 2018. New Orleans, USA

- Ramachandran et al. Unsupervised Pretraining for Sequence to Sequence Learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 383–391 Copenhagen, Denmark, September 7–11, 2017.

Thank you!

Any Questions?

Our MT team is hiring a new NLP Data Scientist!

Contact me if you are interested

Talaat Khalil: talaat.khalil@booking.com