



DIPARTIMENTO
DI INFORMATICA
Università di Pisa

Low Resource Neural Machine Translation

Michele Resta

21/05/2020

Outline

- Introduction
 - Machine Translation Problem
- Neural Machine Translation crash course
 - Sequence to sequence architecture
 - Decoding
 - NMT evaluation
- Low Resource NMT
 - Corpora and domains
 - Learning techniques
- Conclusions

Machine Translation

Machine Translation (MT) is the task of translating a sentence x from one language (the source language) to a sentence y in another language (the target language).

How can we solve this task?

Machine Translation

- Core idea: learn a probabilistic model from data
- Suppose we are interested in translating French \rightarrow English
- We want to find the best target sentence y (English) given the source sentence x French:

$$\operatorname{argmax}_y P(y | x)$$

Machine Translation

What do we need to learn a translation model $P(y|x)$?

1. A **large** amount of parallel data
2. A flexible model architecture

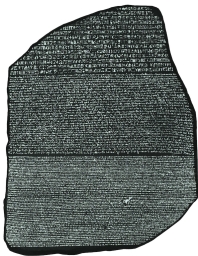


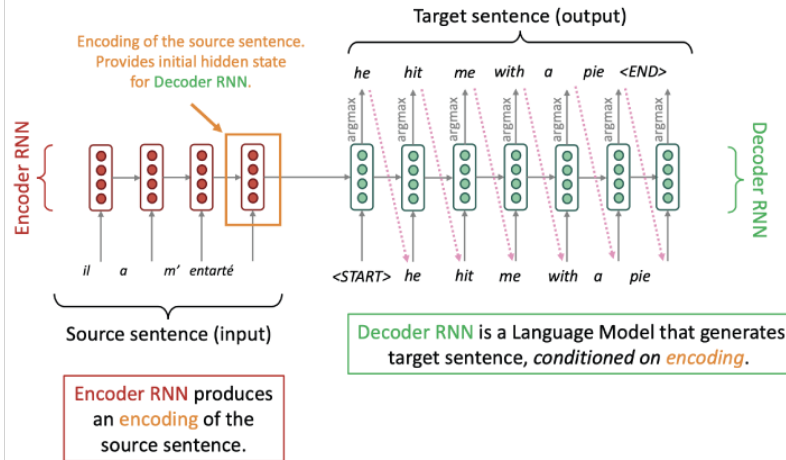
Figure: The Rosetta stone. The same text is written in Egyptian hieroglyphs, Demotic and Ancient Greek

Neural machine Translation

- Neural Machine Translation (NMT) is a way to do Machine Translation with a single end-to-end neural network
- The architecture is called a sequence-to-sequence model (aka seq2seq) and it involves two RNNs

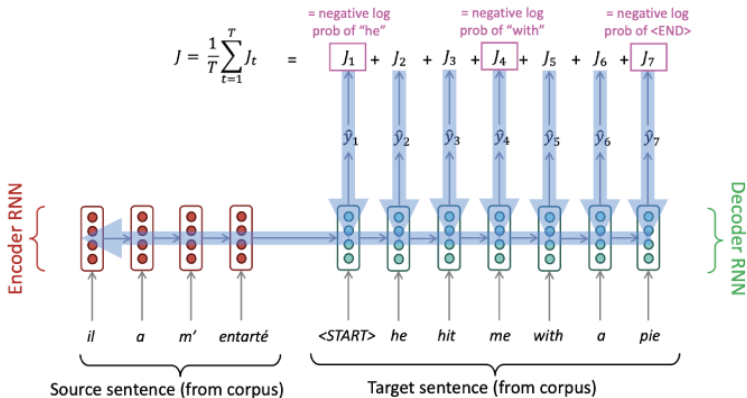
Neural machine Translation

Seq2Seq Architecture



Neural machine Translation

Seq2Seq Training



Neural Machine Translation

- Seq2seq is trained "end-to-end" with backpropagation
- The system calculate the probability of next target word given the source sentence x , and the target words y_i available up to now
- so NMT calculates $P(y | x)$:

$$P(y | x) = P(y_1 | x) P(y_2 | y_1, x) \dots P(y_T | y_1, \dots, y_{T-1}, x)$$

Neural Machine Translation

Decoding

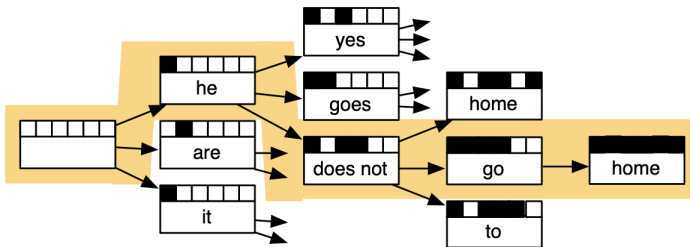
- Ideally we want to find a translation of length T that maximizes

$$P(y | x) = \prod_{t=1}^T P(y_t | y_1, \dots, y_{t-1}, x)$$

- An exhaustive search of the possible sequences y is not feasible: $O(V^T)$ complexity with V = vocabulary size (around 10^5 for bilingual systems)
- Strategies:
 - Greedy decoding
 - Beam Search

NMT Decoding

Beam Search



Idea: At each step of the decoder keep track of the best k partial translations (hypotheses)

NMT Decoding

Beam Search

- k is the beam size (around 10)
- Each hypotheses has a score
- Whenever an hypothesis contains the $\langle \text{END} \rangle$ token that translation is complete
- We stop the search when:
 - We reach a cutoff length T
 - We have at list a predefined number of completed translations

NMT Summary

- Transformers are the most used architecture in NMT
 - Huge feedforward neural network composed of an encoder and decoder (same seq2seq paradigm)
 - Heavy use of attention
- Attention is a general technique to compute a weighted sum of vectors given a query
 - Help the decoder focusing on relevant part of the input sentence
 - Different types of attention exists (scaled, dot, etc..)
 - Dramatically improves NMT systems

NMT Automatic Evaluation

Bilingual Evaluation Understudy

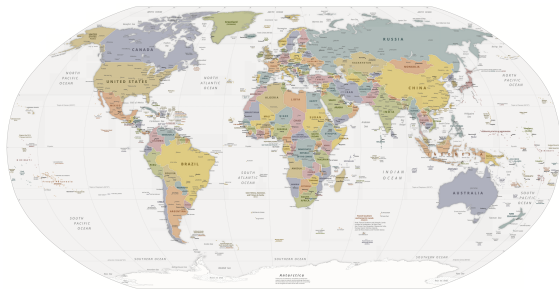
- MT Systems evaluated using BLEU: output translations are compared against one or more human references combining:
 - 1-2-3-4-gram (clipped) precision
 - a penalty factor for short sentences
- Everybody hates BLEU, but everybody uses it (despite its limitation)
 - simple n-gram overlapping
 - no consideration of syntactic structure of the sentences

Outline

- Introduction
 - Machine Translation definition
- Neural Machine Translation Crash-Course
 - Sequence to sequence architecture
 - Decoding
 - NMT evaluation
- Low Resource NMT
 - Corpora and domains
 - Learning techniques
- Conclusions

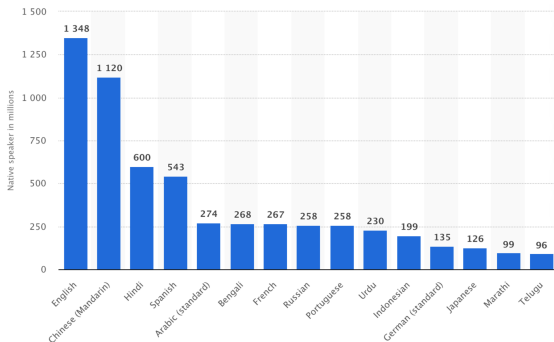
Low Resource NMT

Motivation



- More than 6000 languages in the world
- 80% of the world population don't speak English
- Less than 5% of the population is a native English speaker

Most spoken languages worldwide in 2021

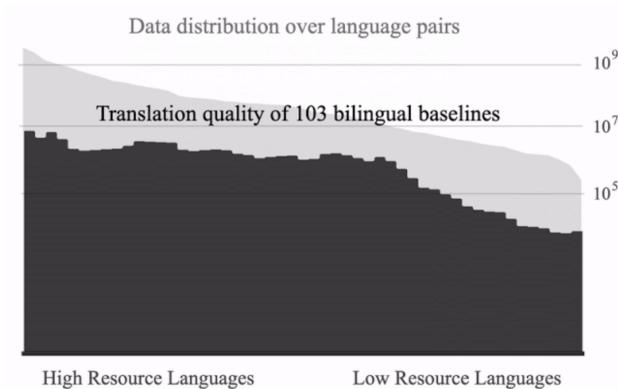


- Top 10 languages spoken by half of the world population
- Remaining languages spoken by the other half
- Lots of languages spoken by less than 1000 people

How much is "Low"?

- A language is considered a low resource language if the amount of parallel sentences is $\leq 10^4 \sim 100\text{KB}$
- Problematic because the size of current NMT systems is in the order of 10^8 parameters
- Google's production dataset sizes for its multilanguage NMT system, are in the order of TB

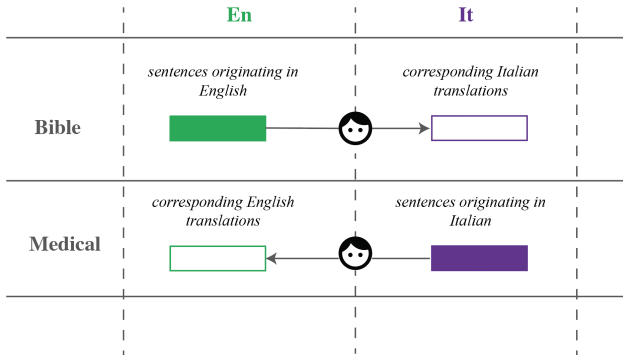
Data influence on translation













The data distribution over all language pairs (in log scale) and the relative translation quality (BLEU score) of the bilingual baselines trained on each one of these specific language pairs.

source: [Google AI blog](#)

Parallel dataset and domains



Low Resource NMT Learning setting

	En	Fr	It	Sw
Bible				
Medical				
News				
Fairy tales				

- Some data may be available only in a single language.
- Some domain are not covered at all in some languages

Challenges

- Data availability
- Domain mismatch
- Quality of data

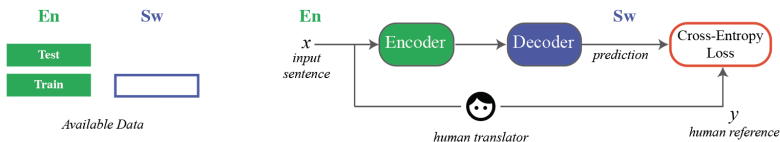
How do we learn in this difficult setting?

Outline

- Introduction
 - Machine Translation definition
- Neural Machine Translation Crash-Course
 - Sequence to sequence architecture
 - Decoding
 - NMT evaluation
- Low Resource NMT
 - Corpora and domains
 - Learning techniques
- Conclusions

Small data available in target language

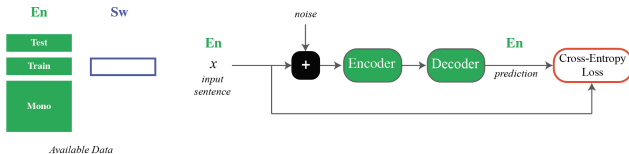
Supervised Learning



- Training dataset $\mathcal{D} = \{(x, y)_i\}_{i=1, \dots, N}$
- Model trained using cross entropy loss (attention-based Transformer)
- If N is small the model will need regularization
 - Dropout

Monolingual corpus in source language

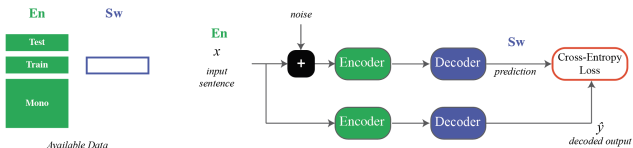
Semi-supervised Learning



- Training dataset $\mathcal{D} = \{(x, y)_i\}_{i=1, \dots, N}$
- Large monolingual corpus $\mathcal{M}^S = \{x_j^S\}_{j=1, \dots, M_S}$
- Try to model $P(x)$ with the encoder
- Loss: $\mathcal{L}^{DAE}(\theta) = -\log P(x | x + n)$
- Pre-train encoder and use it in a supervised system
- Add the Denoising loss to a supervised system

Monolingual corpus in source language

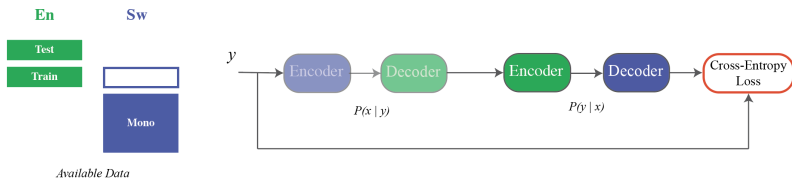
Self Training



1. train a model $P(y | x)$ on \mathcal{D}
2. decode $x \sim \mathcal{M}^S$ to \hat{y} and create a new corpus
$$\mathcal{A}^S = \left\{ \left(x_j^S, \hat{y}_j \right) \right\}_{j=1, \dots, M_S}$$
3. retrain on $\mathcal{D} \cup \mathcal{A}^S$
 - repeat steps 2 and 3 as long as the system improves
 - Loss: $\mathcal{L}(\theta) = \mathcal{L}^{sup}(\theta) - \lambda \log P(\hat{y} | x + n)$

Monolingual corpus in target language

Back-Translation (BT)



1. Monolingual corpus $\mathcal{M}^t = \left\{ x_k^t \right\}_{k=1, \dots, M_t}$
2. Train two models on parallel data \mathcal{D}
3. Decode $y^t \sim \mathcal{M}^t$ to \hat{x} and create $\mathcal{A}^t = \left\{ \left(\hat{x}_k, y_k^t \right) \right\}_{k=1, \dots, M_t}$
4. Retrain on $\mathcal{D} \cup \mathcal{A}^t$

Outline

- Introduction
 - Machine Translation definition
- Neural Machine Translation Crash-Course
 - Sequence to sequence architecture
 - Decoding
 - NMT evaluation
- Low Resource NMT
 - Corpora and domains
 - Learning techniques
- Conclusions

Conclusion

Wrap-up

- Training paradigm has to be adapted according to available data
- iterative BT, denoising pretraining and multilingual training perform pretty well on low resource languages
- Combining the approaches is quite challenging

Conclusion

Open challenges

- Data quality and domain mismatch
- Corpora sizes differences among languages
- Training huge models to exploit larger quantities of data in a multilingual setting

Thanks for your attention!

michele.resta@phd.unipi.it
Room 300 - CS Department, Pisa

References & Acknowledgments

- [1] Sutskever et al. *Sequence to Sequence Learning with Neural Networks*.
- [2] Vaswani et al. *Attention is all you need*.
- [3] He et al. *Revisiting Self-Training for Neural Sequence Generation*. 2017
- [4] Liu et al. *Multilingual Denoising Pre-training for Neural Machine Translation*.
- [5] Sennrich et al. *Improving Neural Machine Translation Models with Monolingual Data*. CoRR 2015
 - Schemes of slides 6 and 7 from Christopher Manning's NLP course