Metrics and Models– L5

Bill Byrne

Lent 2022

Neural Machine Translation and Dialogue Systems – MLMI8

MPhil in Machine Learning and Machine Intelligence

UNIVERSITY OF CAMBRIDGE

Overview

Evaluation

- Translation and Dialogue
  - Manual and automatic evaluations

Modelling

- Subwords
- PreTraining and Fine-Tuning
  - Model Architectures:
    - BERT and GPT-2
  - Pre-Trained/Fine-Tuned model applications:
    - Dialogue
    - Automatic Metrics
  - Inline and embedded tags
    - Multilingual Neural Machine Translation

Evaluation

Translation and Dialogue, Human and Automatic, Fluency and Accuracy

Evaluating Machine Translation – by Humans

Machine translation quality is determined by the task it is expected to accomplish.

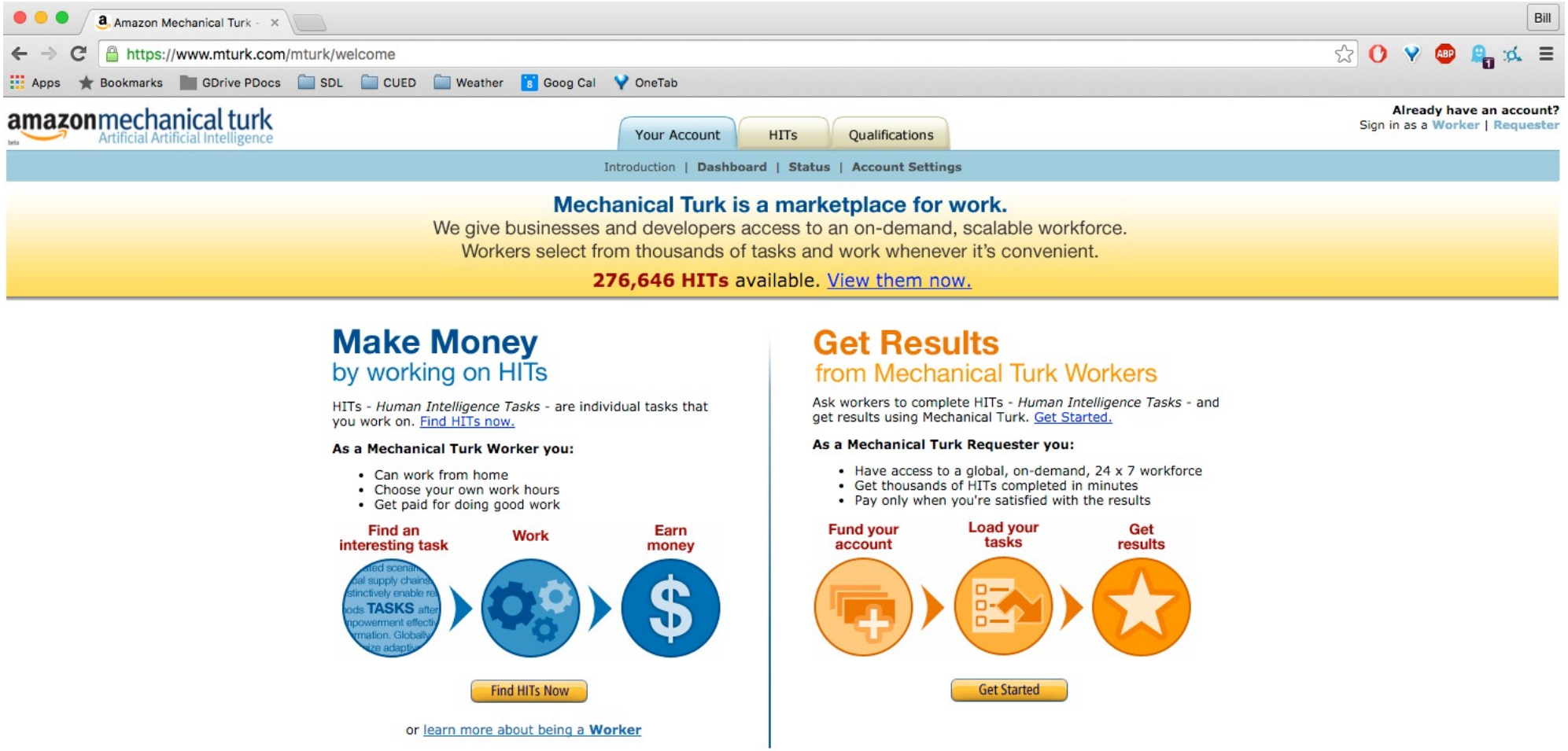
For a given task, humans must decide what is a 'good' or a 'bad' translation.

Human Evaluation scenarios --

- Human judges assess translations and provide:
  - Binary scores: correctness, task completion
  - Scores on a scale: fluency, adequacy (e.g. on a range from 1 to 5)
  - Preference scores: which is better?
  - Post-edit scores: minimal output correction ← direct commercial impact

× Human evaluation is slow and costly → cannot be used alone for system development

Mechanical Turk

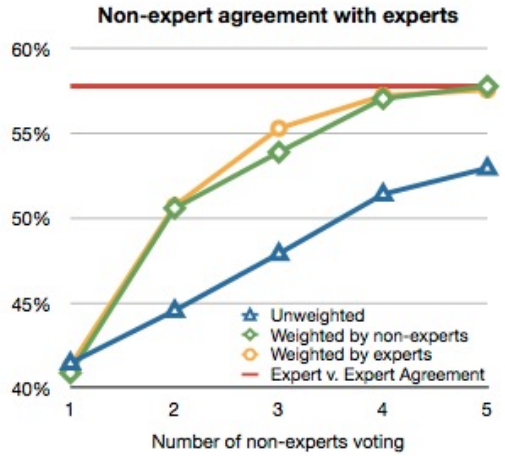


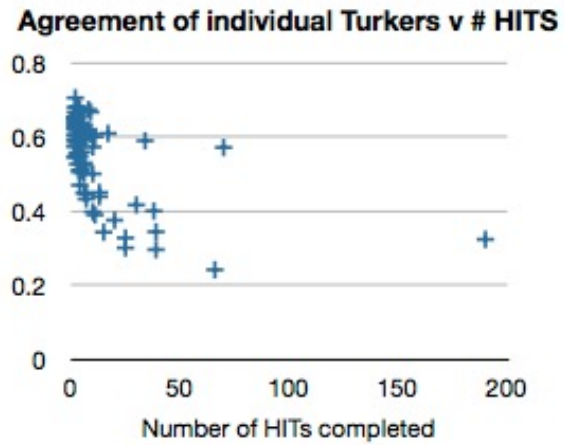
Snow et al. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks" EMNLP'08

Mechanical Turk for MT (and Dialogue) Evaluation

Instructions given to the Turkers:

- ***Evaluate machine translation quality*** *Rank each translation from Best to Worst relative to the other choices (ties are allowed).*
- *If you do not know the source language then you can read the reference translation, which was created by a professional human translator*





Non-expert annotators can rank translation quality, but judgements are still expensive...

Automatic Metrics for Machine Translation

BLEU is a metric based on n-gram precision against a set of reference translations *R*

- A single metric, meant to capture both fluency and adequacy
  - N is the maximum n-gram order (N=4)
  - p_n(T,R) is a *clipped* n-gram precision for translation *T*
  - γ(T, R) is a *brevity penalty* that penalises translations that are shorter than the references

$$BLEU(T,R) = \gamma(T,R)\exp\left(\sum_{n=1}^{N}\frac{1}{N}\log p_n(T,R) \times 100\right)$$

Reference : mr. speaker , in absolutely no way .
Hypothesis : in absolutely no way , mr. chairman .

BLEU Computation

n-gram matches				BLEU
1-word	2-word	3-word	4-word	$\left(\frac{7}{8}\times\frac{3}{7}\times\frac{2}{6}\times\frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$
7/8	3/7	2/6	1/5	

UNIVERSITY OF CAMBRIDGE

Papineni, K., et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation.

Translationese and BackTranslation

Automatic metrics such as BLEU are no better than the data on which they are based

Translated texts in a human language exhibit unique characteristics that set them apart from texts originally written in that language.

- Awkwardness or ungrammaticality of translation, such as due to overly literal translation of idioms or syntax
- Compared to original texts, translations tend to be simpler, more standardised, and more explicit and they retain some characteristics that pertain to the source language.

Effects on training and evaluation:

- Translationese is artificially easier to translate, resulting in inflated scores for MT systems.
- MT systems perform better when trained on parallel data whose source side is original and whose target side is translationese

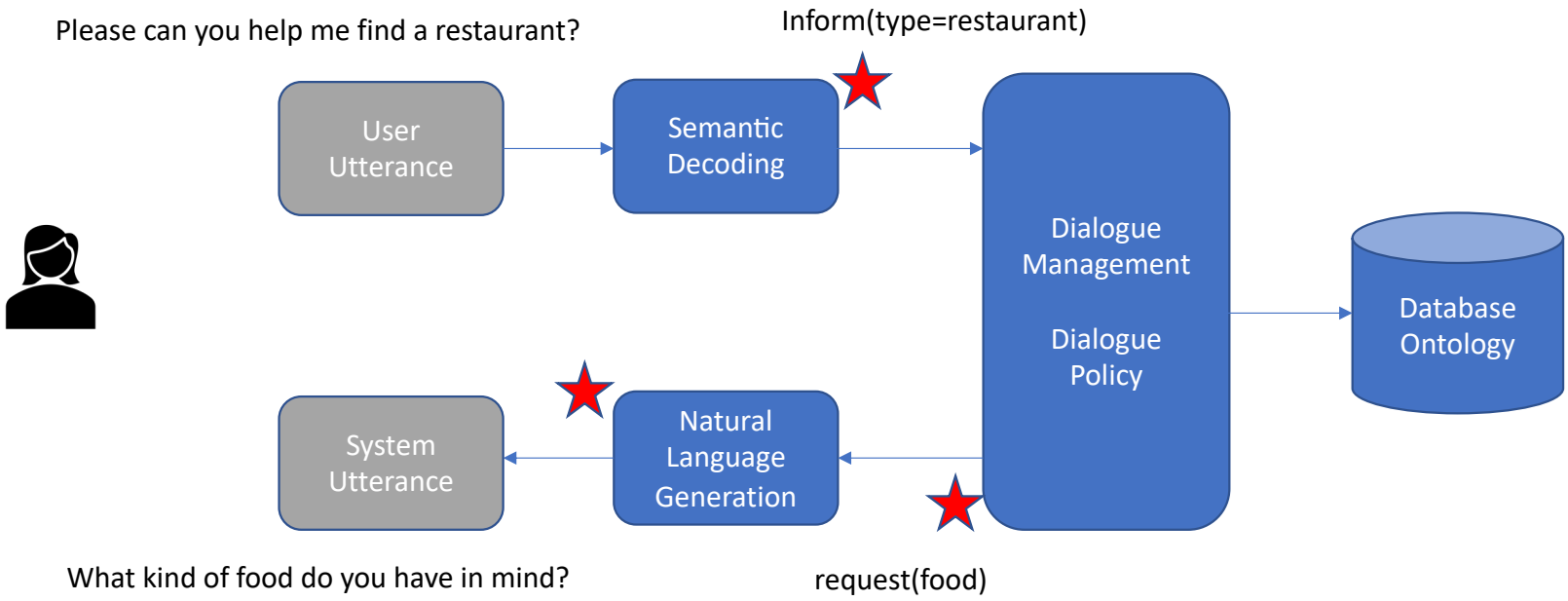
Effects are complicated when systems are trained with **backtranslated** parallel text

- Zhang et al. The Effect of Translationese in Machine Translation Test Sets. WMT'19
- Toral et al. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. WMT'18
- Sennrich et al. Improving Neural Machine Translation Models with Monolingual Data. ACL'16

Evaluating Task Oriented Dialogues

- For task-based dialogue, if the task is unambiguous, we can simply measure absolute task success
  - ... did the system book the correct plane flight, or put the correct event on the calendar, ...

- If we have annotated dialogues, we can measure the performance of individual system components
  - Annotated corpora are good for system development, but can be expensive and can be limited in diversity



Evaluating Task Oriented Dialogue System Components Against Annotated Dialogues

Dialogue State Tracking and Dialogue Policy

- Performance can be measured using typical retrieval metrics:
  - Error rate, precision, recall, F1
- Can be measured at the turn level, if turn level annotations are available

- A Slot Error Rate example:

  Make an appointment with Chris at 10:30 in Gates 104

Slot	Filler
PERSON	Chris
TIME	11:30 a.m.
ROOM	Gates 104

$$\text{Slot Error Rate for a Sentence} = \frac{\text{\# of inserted/deleted/subsituted slots}}{\text{\# of total reference slots for sentence}}$$

  - Slot error rate is 1/3, since TIME is wrong

Evaluating Task Oriented Dialogue System Components Against Annotated Dialogues

Natural Language Generation

BLEU scores (and other metrics) with dialogue transcriptions as references

- As with translation, the metrics should allow for multiple good answers

Fung et al. 2020

Evaluating Task Oriented Dialogue System Components Against Annotated Dialogues

System Level Performance

Automatic metrics, relative to the original dialogue goals:

- *Inform:* whether the system has provided an appropriate entity
- *Success:* whether all requested attributes have been supplied
- *Fluency:* measured by BLEU score

	Baseline	GPT	GPT2-S	GPT2-M
Inform (%)	**76.7**	71.53	66.43	70.96
Success (%)	**64.63**	55.36	55.16	61.36
BLEU (%)	18.05	17.80	18.02	**19.05**

Table 1: Evaluation on MultiWOZ with the greedy sampling procedure.

Human Evaluation

- Mechanical Turk (or similar) can be used to collect human judgements
  - absolute quality metrics
  - pair-wise system comparisons, at the turn level or dialogue level
    - *The turkers were required to choose what response they prefer when presented with two responses from two different models, resulting in more than 300 scores per each model pair.*

Model 1	vs		Model 2
GPT	59 %	41%	Baseline
GPT	46 %	54 %	Target
GPT2	46 %	54 %	Target
GPT2	45 %	55 %	Baseline
Baseline	43 %	57 %	Target
GPT2	51 %	49 %	GPT

Table 3: Human ranking of responses between all pairs of four analyzed models and the original responses.

Budzianowski, Vulic, *Hello, It's GPT-2 - How Can I Help You?* Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems https://arxiv.org/abs/1907.05774

Task Oriented Dialogue System Development, without turn level annotation

- In realistic dialogue scenarios, success or failure is not known until the dialogue is completed
- To decide which action to take, a reinforcement learning system gets a reward at the end of the dialogue, and uses that reward to train a policy to take actions.
- Less emphasis recently on these techniques
  - user simulators
  - dialogue policy optimisation

Policy deterministic $\pi : \mathcal{B} \rightarrow \mathcal{A}$ or stochastic $\pi : \mathcal{B} \times \mathcal{A} \rightarrow [0, 1]$

Return $R_t = \sum_{k=0}^{T-l} \gamma^k r_{t+k}$

Value function How good is it for the system to be in a particular belief state?

$$V_\pi(\mathbf{b}) = E_\pi \left[\sum_{k=0}^{T-t} \gamma^k r_{t+k} | b_t = \mathbf{b} \right]$$

Q-function What is the value of taking action a in belief state $\mathbf{b}$ under a policy π?

$$Q_\pi(\mathbf{b}, a) = E_\pi \left[\sum_{k=0}^{T-t} \gamma^k r_{t+k} | b_t = \mathbf{b}, a_t = a \right]$$

Levin, Pieraccini, Eckert *A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies.* 2000

Modelling

- BERT, GPT-2, Pre-Training and Fine-Tuning
- Metrics, Task Oriented and Retrieval Based Dialogue, Multilingual NMT

Subword Units – Byte Pair Encoding

The Input and Output vocabularies of neural models are typically limited to 30-50K words

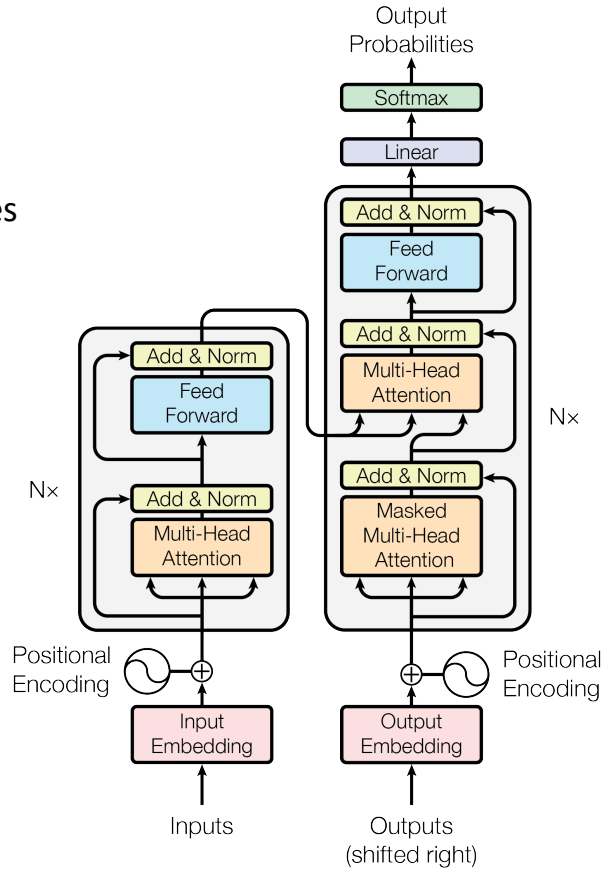
- operational constraints: size of the input and output layers, including the softmax, increases with the vocabulary size
- modelling: words that appear infrequently are not well modelled
- words not seen are Out Of Vocabulary

Subword modelling replaces words by automatically learned subsequences

- controls vocabulary size : longer input and output sequences, but a smaller vocabulary
- individual tokens occur more frequently
- Remaining OOVs arise mainly due to unseen characters

```
(l o w </w>) (l o w e r </w>) (s l o w e r </w>)
(lo w </w>) (lo w e r </w>) (s lo w e r </w>)
(low </w>) (low e r </w>) (s low e r </w>)
(low </w>) (low er </w>) (s low er </w>)
(low </w>) (low er</w>) (s low er</w>)
(low </w>) (lower</w>) (s lower</w>)
```

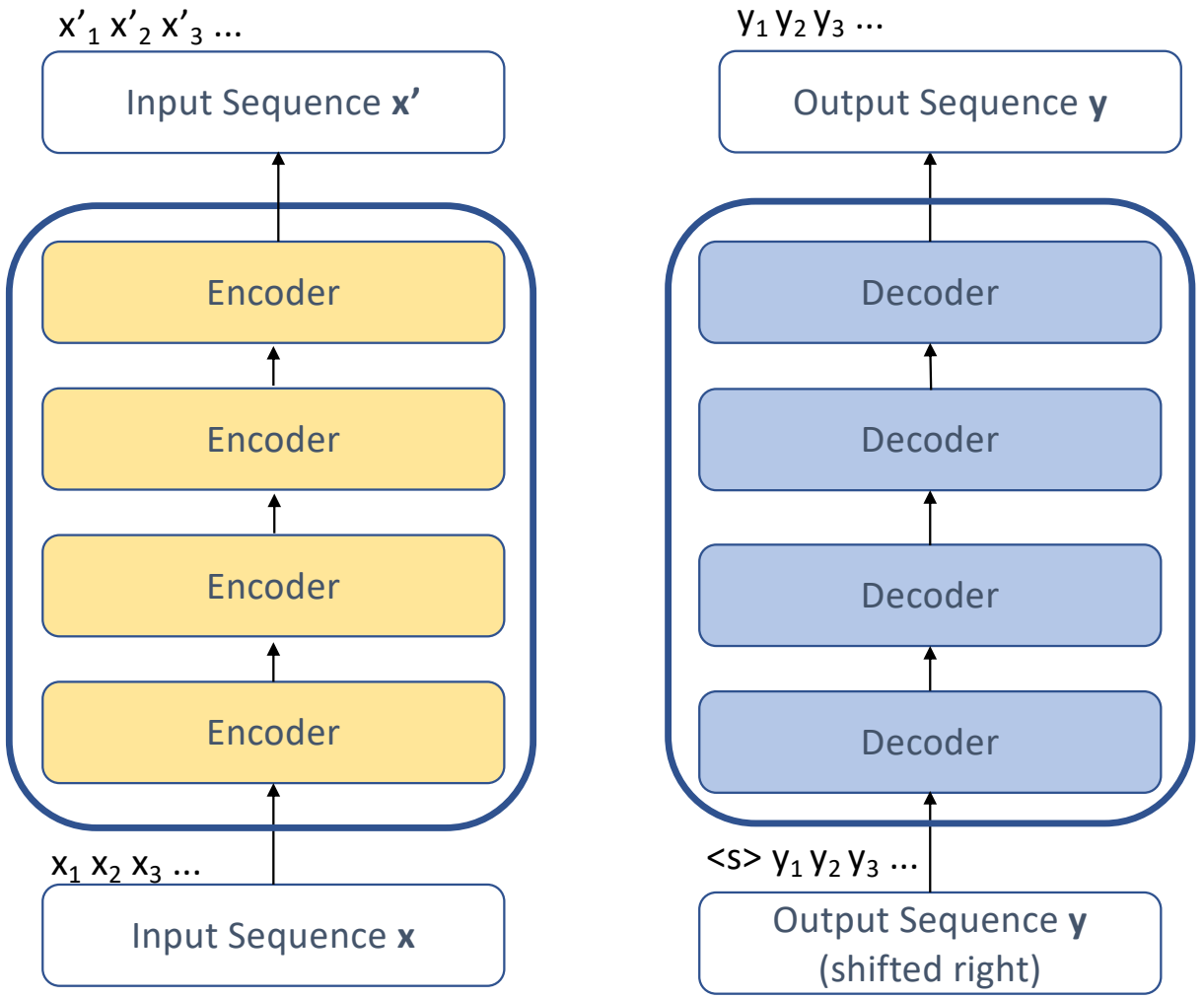
```
l o --> lo
lo w --> low
e r --> er
er </w> --> er</w>
low er</w> --> lower</w>
```



Can be applied using WFSTs

Sennrich, R., Haddow, B., Birch, A. (2015). Neural machine translation of rare words with subword units.
J Gao https://www.mlmi.eng.cam.ac.uk/files/industry_day_poster_j_gao.pdf

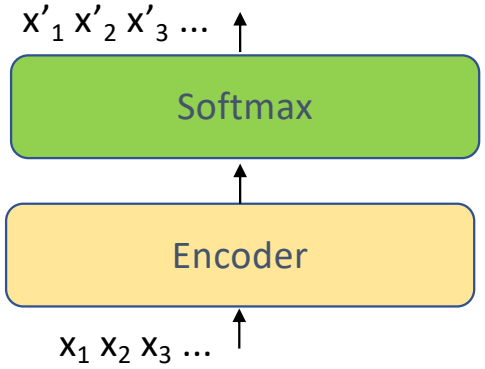
Encoder and Decoder can be used separately



BERT: Bidirectional Encoder Representations from Transformers

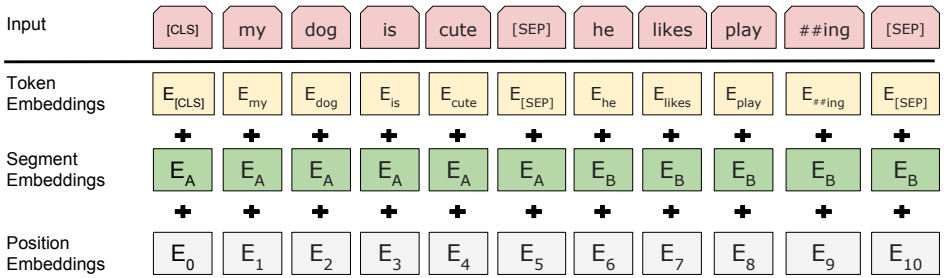
Encoder architecture, with an output softmax layer

- Reads the entire input sequence at once (non-causal)
- Captures the relationship between two sentences



Pre-Training tasks:

- Masked Language Modeling:
  - Randomly replace input sequence tokens by [MASK]
  - Learn to predict the missing item
- Next Sentence Prediction: predict whether two sentences naturally follow each other
  - [CLS] representation is fed to an output layer, for classification



Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

BERTScore: Evaluating Text Generation with BERT

- BLEU: n-gram scores using exact symbol matches
- BERTScore: sequence similarity from sequence embeddings

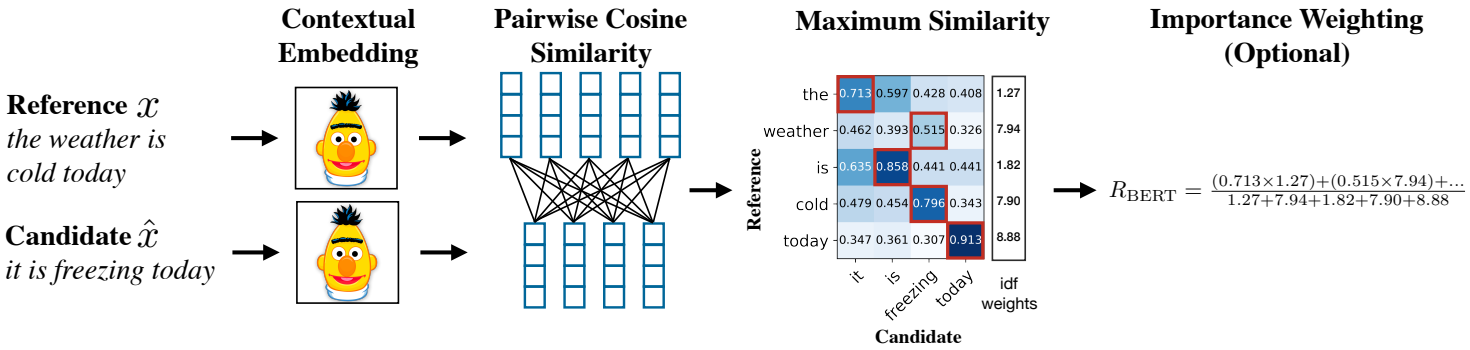


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad .$$

ConveRT: Efficient and Accurate **Conv**ersational **R**epresentations from **T**ransformers

Response Selection is a task of selecting the most appropriate response given the dialog history

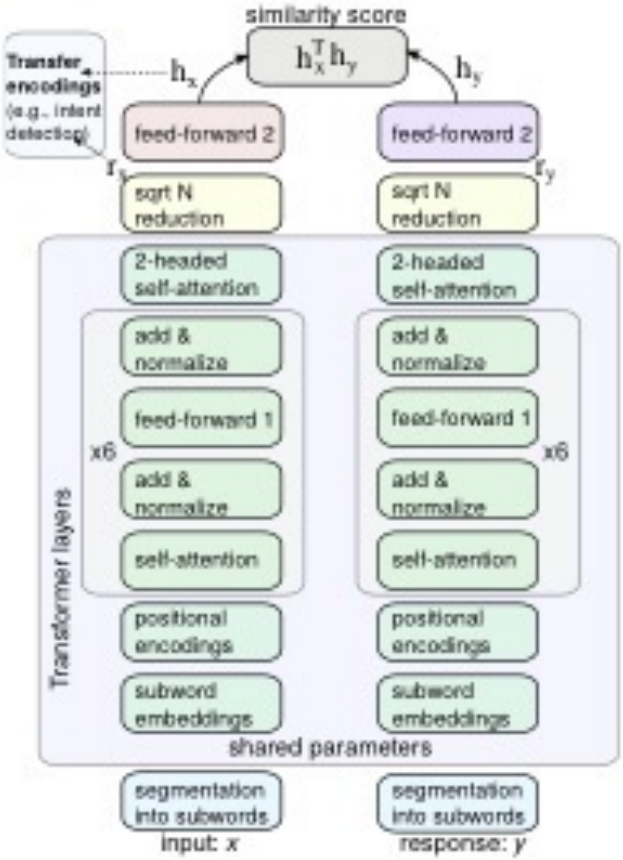
- Central to retrieval-based dialog systems
- Encode the *context* and a large collection of responses in a joint semantic space

Student: I'm very interested in representation learning.
Teacher: Do you have any experience in PyTorch?
Student: Not really.
Teacher: And what about TensorFlow?

- Retrieve a relevant response by matching the query against the encodings of each candidate response.

Pre-training on Reddit data

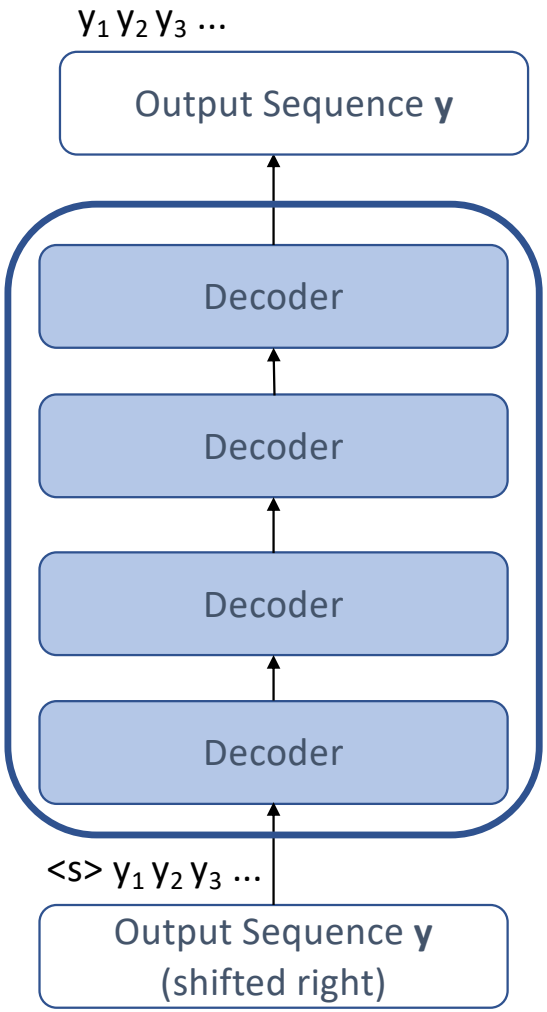
Fine-tuning on task-specific data, e.g. AMAZONQA



GPT-2: Generative Pre-trained Transformer

Features:

- Training data: WebText, outgoing high karma links from Reddit
  - 8 million documents, 40GB text (less Wikipedia)
- Byte Pair Encoding
  - Vocabulary size of 50K
- Slightly modified transformer decoder layer
  - Layer normalisation at the input of each sub-block
  - Additional layer normalisation after the final self-attention block
- Context size of 1024 tokens
- Training with large batch sizes of 512



Radford et al. Language Models are Unsupervised Multitask Learners https://openai.com/blog/better-language-models/

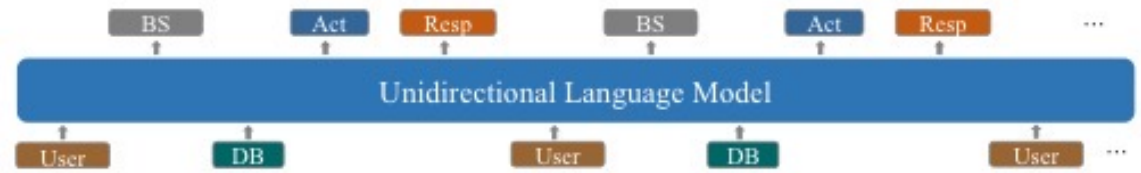
User Utterance, Belief States, DB results, System Acts, System Response

... fine-tune GPT-2 on the sequence of the entire dialog session consisting of user utterance, belief state, database result, system act, and system response of every dialog turn.

- a single training sequence for a dialog with *T* turns: $\{U_0, B_0, D_0, A_0, R_0, ..., U_T, B_T, D_T, A_T, R_T\}$

- ... plain text, delexicalization, tags mark type boundaries, e.g. <sos_u> ... <eos_u>

... It is also [value_price], and rated [value_stars] stars. <eos_r> <sos_u> great. can you book that 1 for 4 nights checking in on tuesday? there will be 6 people. <eos_u> <sos_b> [hotel] internet yes type guest house parking yes stars 4 name alexander b&d and breakfast stay 4 day tuesday people 6 <eos_b> <sos_db> <db_1> <eos_db> <sos_a> [hotel] [offerbooked] reference [general] [reqmore] <eos_a> <sos_r> o



Multilingual Neural Machine Translation

If we're interested in N languages, is it necessary to build Nx(N-1) translation systems?

Can high-resource language pairs help low resource translation?

Build a single Transformer system, and use *tags* to indicate the target language:

- Consider the following English→Spanish pair of sentences:

  Hello, how are you? -> Hola, ¿cómo estás?

  Introduce a tag into the source language:

  <2es> Hello, how are you? -> Hola, ¿cómo estás?

- Many options: N-to-M, 1-to-N, N-to-1, ... possibly grouping by language family

Models can capture complex behavior, e.g. multiple typographic scripts, codeswitching,...

- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

- Johnson et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2017
- Fan et al. Beyond English-Centric Multilingual Machine Translation. 2020

Summary & Key Points

- Current models are so powerful that relatively minor biases in the data used for training and evaluation can have surprisingly large impact
- Pre-trained models can be applied for sentence similarity (BERT) or prediction (GPT-2)
- Models are capable of learning text mixed with annotations
- Metrics can be extended to use continuous representations
- Tags and annotations can control model behaviour
  - as embeddings (BERT) or as inline tags (to specify target language in multilingual NMT)
- There is a convergence in modelling techniques and metrics for NMT and Dialogue Systems