Advanced NLP tasks

Contents

Named Entity Recognition (NER)

NER

Named entity recognition (NER), aims at identifying real-world entity mentions from texts, and classifying them into predefined types.

Gold Dataset

Suxamethonium infusion rate and observed fasciculations.

Suxamethonium chloride (Sch) was administred i.v.

NER

We wish to predict an output vector $\mathbf{y} = (y_1, y_1, \ldots, y_L)$, of random variables, given an observed characteristic vector $\mathbf{x} = (x_1, x_2, \ldots, x_L)$

$\mathbf{y}$ takes it value from a list of N possible values.

Part-of-Speech Tagging (POS)

POS is the process of mapping words in a text with a label corresponding to their grammatical class.

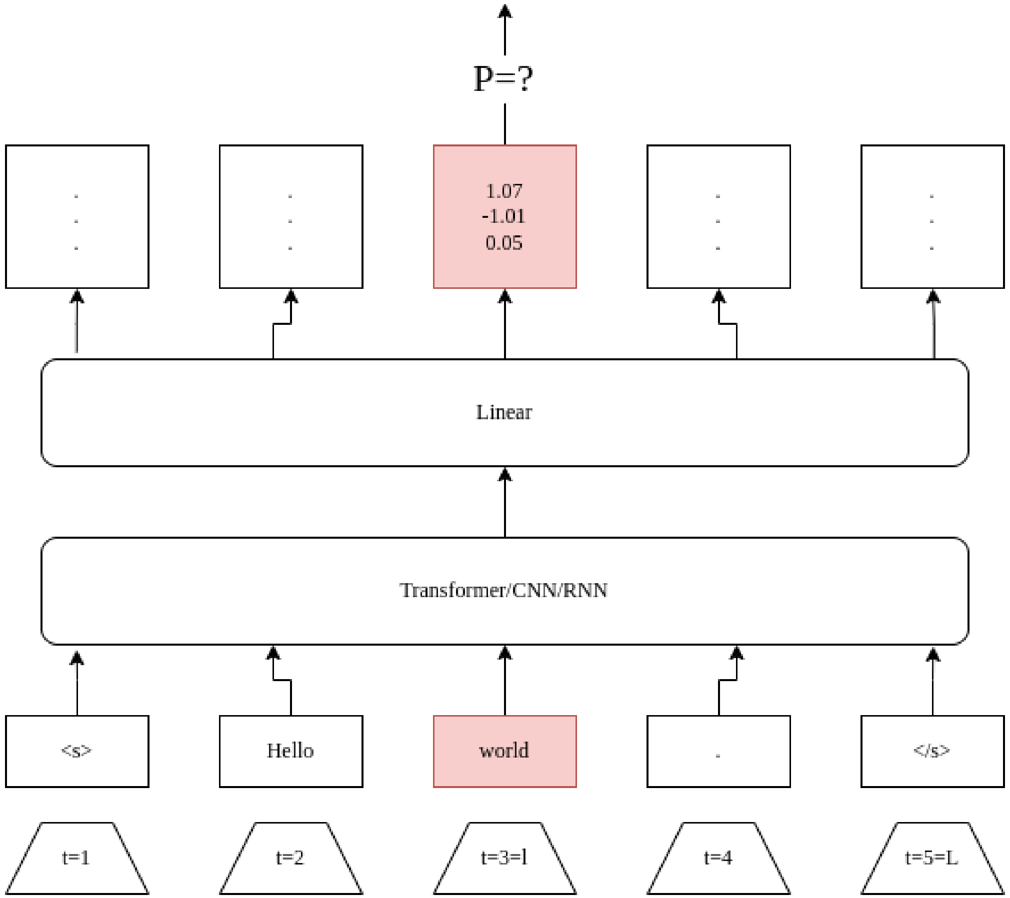
("He", "likes", "to", "drink", "tea"), $\rightarrow$ ("PERSONAL PRONOUN", "VERB", "TO", "VERB", "NOUN").

Part-of-Speech Tagging (POS)

There several levels of granularity.: using the tag set for english

("He", "likes", "to", "drink", "tea"), $\rightarrow$ ("PRP", "VBP", "TO", "VB", "NN").

Conditional Random Field (CRF)



Conditional Random Field (CRF)

For each token in a sentence at position l we want to compute a probability p to belong to a class n.

$$p : f(\mathbf{x}, \theta)_l \mapsto ?$$

with $p \in [0, 1]$

Conditional Random Field (CRF)

Using the softmax function?

$$p : f(\mathbf{x}, \theta)_l^{\mapsto} \frac{e^{f(\mathbf{x},\theta)_l^{(n)}}}{\sum_{n'=1}^{N} e^{f(\mathbf{x},\theta)_l^{(n')}}}$$

The probability given by the softmax function will not encode non-local dependencies!

Conditional Random Field (CRF)

We need to take sequential decisions: what if we add transition scores into our softmax?

$$p : f(\mathbf{x}, \theta)_l \mapsto \frac{e^{f(\mathbf{x},\theta)_l^{(n)} + t(y_l^{(n)}, y_{l-1})}}{\sum_{n'=1}^{N} e^{f(\mathbf{x},\theta)_l^{(n')} + t(y_l^{(n')}, y_{l-1})}}$$

But this is the probability for one token to belong to a class, we want to compute the probability of a whole sequence of label at once...

Conditional Random Field (CRF)

$$P(\mathbf{y}|\mathbf{x}) = \prod_{l=2}^{L} p(\mathbf{y}|f(\mathbf{x},\theta)_l)$$

$$= \prod_{l=2}^{L} \frac{e^{f(\mathbf{x},\theta)_l^{(n)} + t(y_l^{(n)}, y_{l-1})}}{\sum_{n'=1}^{N} e^{f(\mathbf{x},\theta)_l^{(n')} + t(y_l^{(n')}, y_{l-1})}}$$

$$P(\mathbf{y}|\mathbf{x}) = \frac{exp[\sum_{l=2}^{L}\left(f(\mathbf{x},\theta)_l^{(n)} + t(y_l^{(n)}, y_{l-1})\right)]}{\sum_{n'=1}^{N} exp[\sum_{l=2}^{L}\left(f(\mathbf{x},\theta)_l^{(n')} + t(y_l^{(n')}, y_{l-1})\right)]}$$

$$= \frac{exp[\sum_{l=2}^{L}\left(U(\mathbf{x}, y_l^{(n)}) + T(y_l^{(n)}, y_{l-1})\right)]}{\sum_{n'=1}^{N} exp[\sum_{l=2}^{L}\left(U(\mathbf{x}, y_l^{(n')}) + T(y_l^{(n')}, y_{l-1})\right)]}$$

$$= \frac{exp[\sum_{l=2}^{L}\left(U(\mathbf{x}, y_l^{(n)}) + T(y_l^{(n)}, y_{l-1})\right)]}{Z(\mathbf{x})}$$

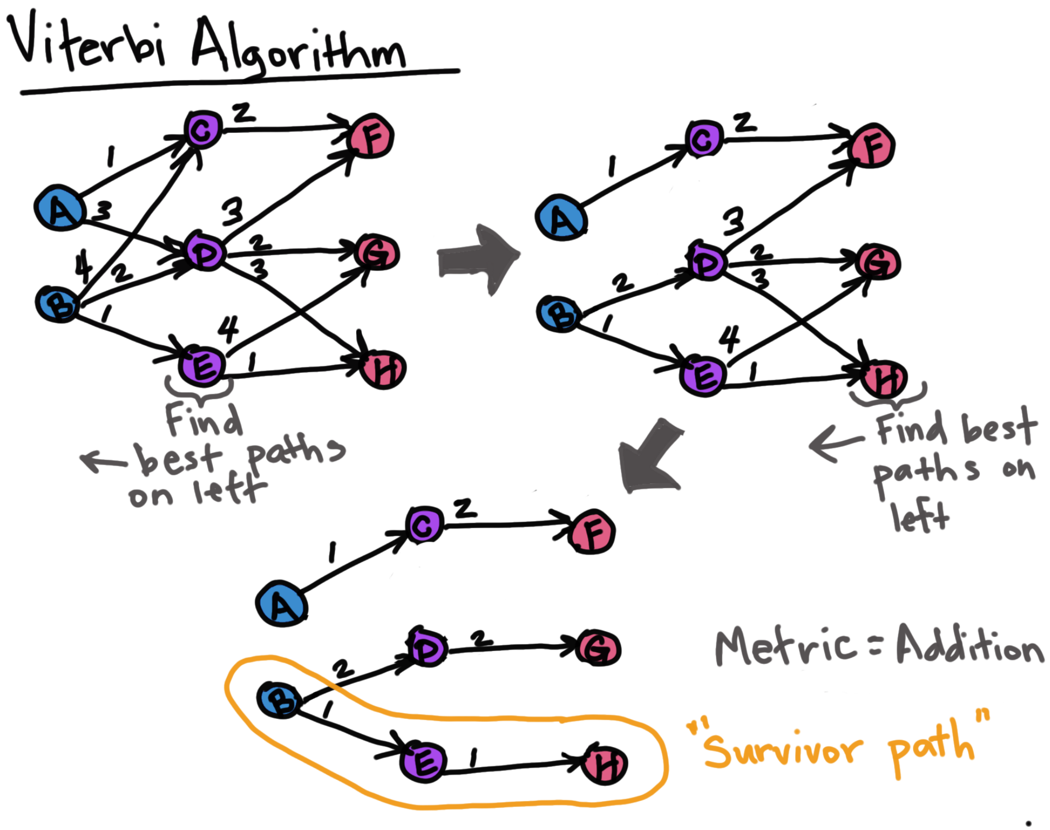
Conditional Random Field (CRF)

$Z(\mathbf{x})$ is commonly referred as the partition function. However, its not trivial to compute: we'll end up with a complexity of $\mathcal{O}(N^L)$.

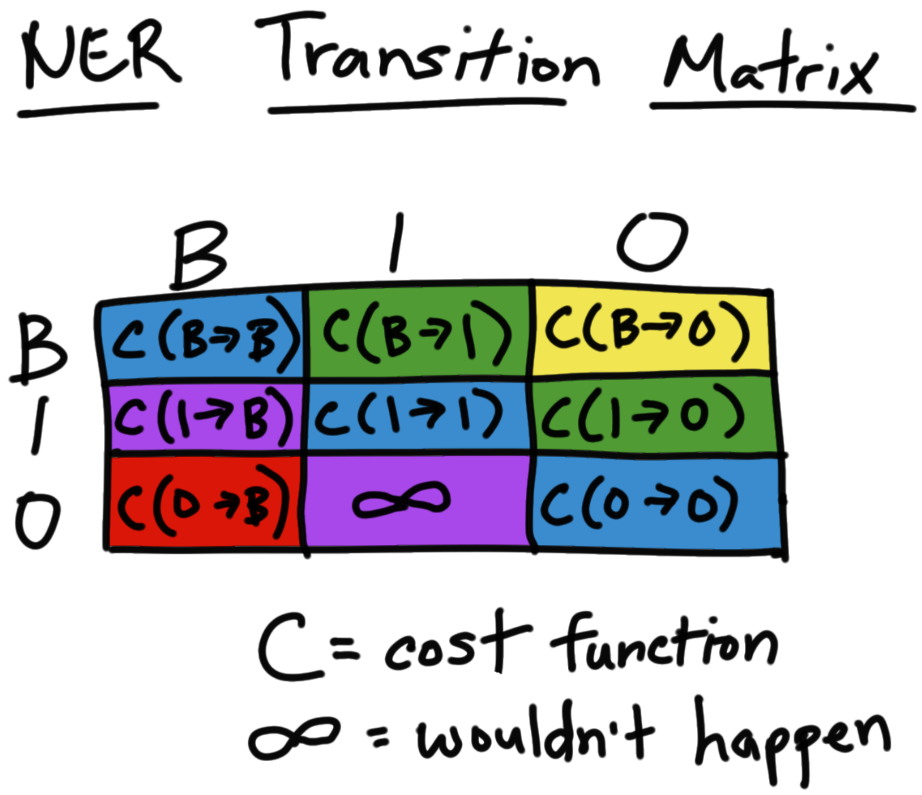
Where N is the number of possible labels and L the sequence length.

How do we proceed?

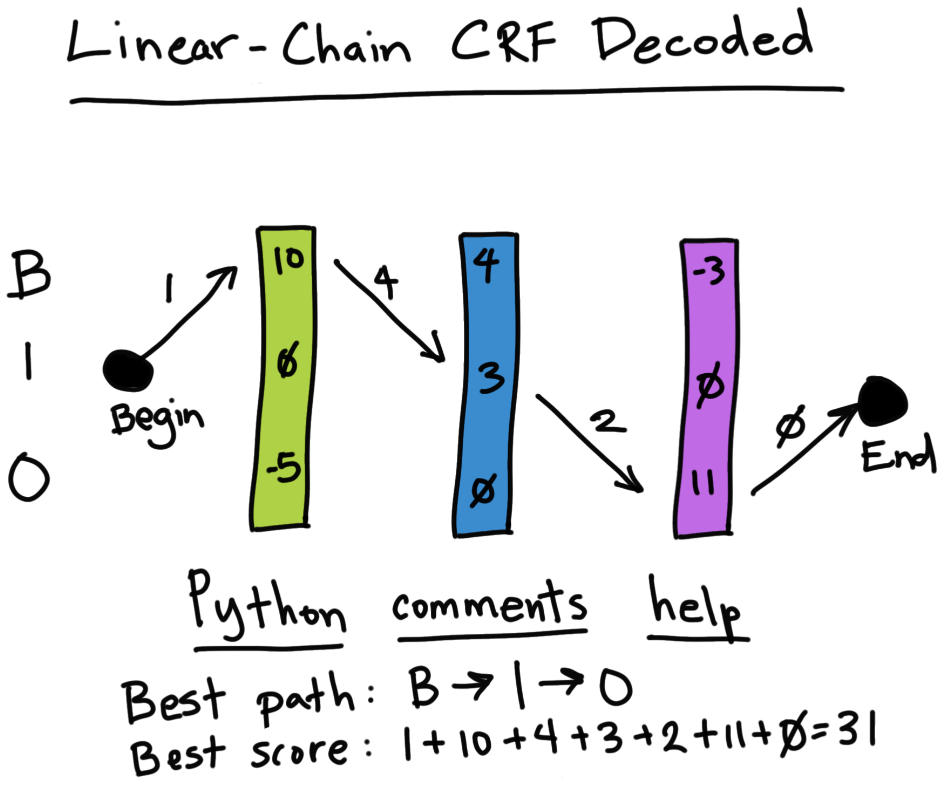
Conditional Random Field (CRF)



Conditional Random Field (CRF)



Conditional Random Field (CRF)



Conditional Random Field (CRF)

Negative log-likelihood:

$$
\begin{aligned}
\mathcal{L} &= -log(P(\mathbf{y}|\mathbf{x})) \\
&= -log(\frac{exp[\sum_{l=2}^{L}(U(\mathbf{x}, y_l^{(n)}) + T(y_l^{(n)}, y_{l-1}))]}{Z(\mathbf{x})}) \\
&= -[log(exp[\sum_{l=2}^{L}(U(\mathbf{x}, y_l^{(n)}) + T(y_l^{(n)}, y_{l-1}))]) - log(Z(\mathbf{x}))] \\
&= log(Z(\mathbf{x})) - \sum_{l=2}^{L}(U(\mathbf{x}, y_l^{(n)}) + T(y_l^{(n)}, y_{l-1}))
\end{aligned}
$$

Conditional Random Field (CRF)

There is an effective way to compute $log(Z(\mathbf{x}))$ with a complexity of $\mathcal{O}(L)$ using the Log-Sum-Exp trick.

$$log(Z(\mathbf{x})) = log(\sum_{n'=1}^{N} exp[\sum_{l=2}^{L} (U(\mathbf{x}, y_l^{(n')}) + T(y_l^{(n')}, y_{l-1}))])$$

$$= c + log(\sum_{n'=1}^{N} exp[\sum_{l=2}^{L} (U(\mathbf{x}, y_l^{(n')}) + T(y_l^{(n')}, y_{l-1})) - c])$$

Conditional Random Field (CRF)

If we fix
$c = max\{U(\mathbf{x}, y_l^{(1)}) + T(y_l^{(1)}, y_{l-1}), \ldots, U(\mathbf{x}, y_l^{(N)}) + T(y_l^{(N)}, y_{l-1})\}$
we ensure that the largest positive exponentiated term is $exp(0) = 1$.

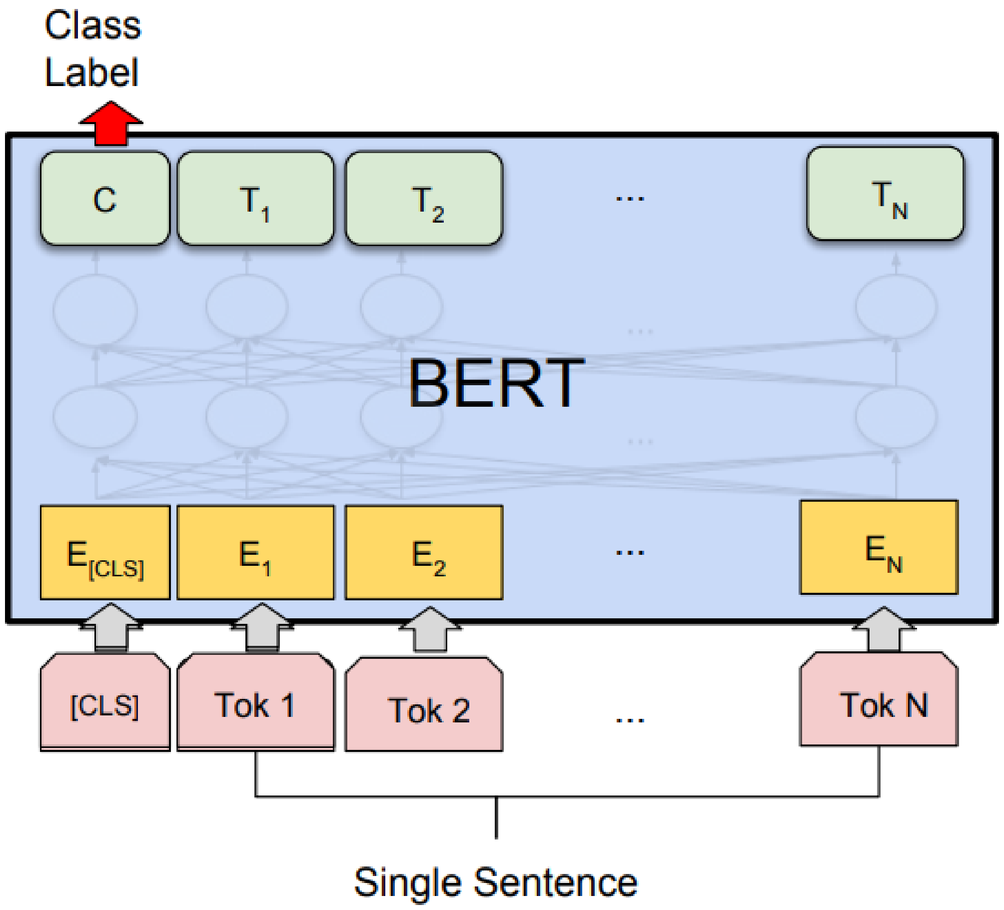
Sentiment Analysis

Sentiment Analysis

Sentiment analysis is a sentence classification task aiming at **automatically mapping data to their sentiment**.

It can be **binary** classification (e.g., positive or negative) or **multiclass** (e.g., enthusiasm, anger, etc)

Sentiment Analysis



Sentiment Analysis

The loss can be the likes of cross-entropy (CE), binary cross-entropy (BCE) or KL-Divergence (KL).

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n'=1}^{N} y^{(n)}.\,log(f(\mathbf{x},\theta)^{(n)})$$

$$\mathcal{L}_{BCE} = -y^{(n)}.\,log(f(\mathbf{x},\theta)^{(n)}) + (1 - y^{(n)}).\,(1 - f(\mathbf{x},\theta)^{(n)})$$

$$\mathcal{L}_{KL} = -\frac{1}{N}\sum_{n'=1}^{N} y^{(n)}.\,log(\frac{y^{(n)}}{f(\mathbf{x},\theta)^{(n)}})$$

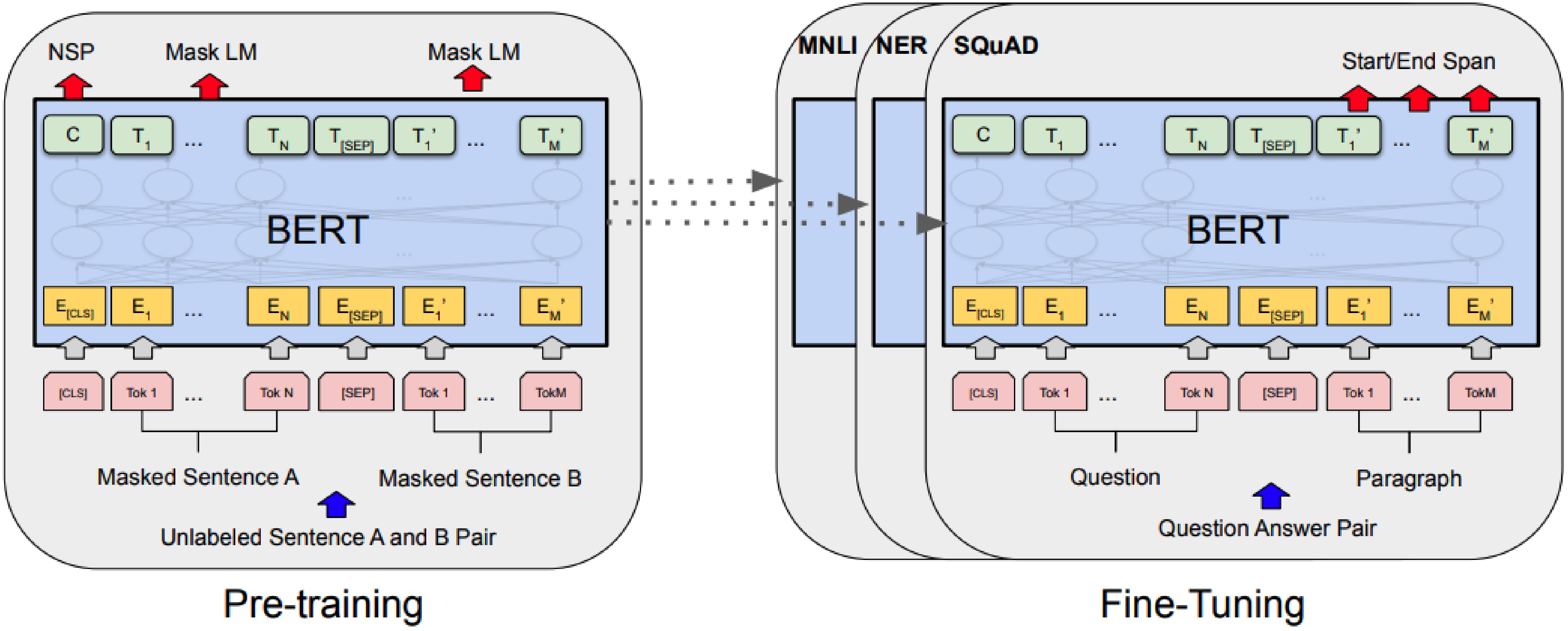
Question Answering (QA)

Question Answering (QA)

QA is the task of **retrieving a span of text from a context** that is best suited to answer a question.

This task is extractive -> **information retrieval**

Question Answering (QA)



Question Answering (QA)

Start End

Token	Start	End
[CLS]	0.	0.
This	0.	0.
is	0.	0.
the	0.	0.
question	0.	0.
[SEP]	0.	0.
This	0.	0.
is	0.	0.
the	0.	0.
context	0.	0.
with	0.	0.
lots	0.	0.
of	0.	0.
info	0.	0.
##rma	0.	0.
##tion	0.	0.
.	0.	0.
Some	0.	0.
use	0.	0.
##less	0.	0.
.	0.	0.
The	1.	0.
answer	0.	0.
is	0.	0.
here	0.	1.
some	0.	0.
more	0.	0.
words	0.	0.
.	0.	0.
[SEP]	0.	0.

Question Answering (QA)

The loss is the cross entropy over the output of the starting token and the ending one:

$$\mathcal{L}_{CE_{QA}} = \mathcal{L}_{CE_{start}} + \mathcal{L}_{CE_{end}}$$

Natural Language Inference (NLI)

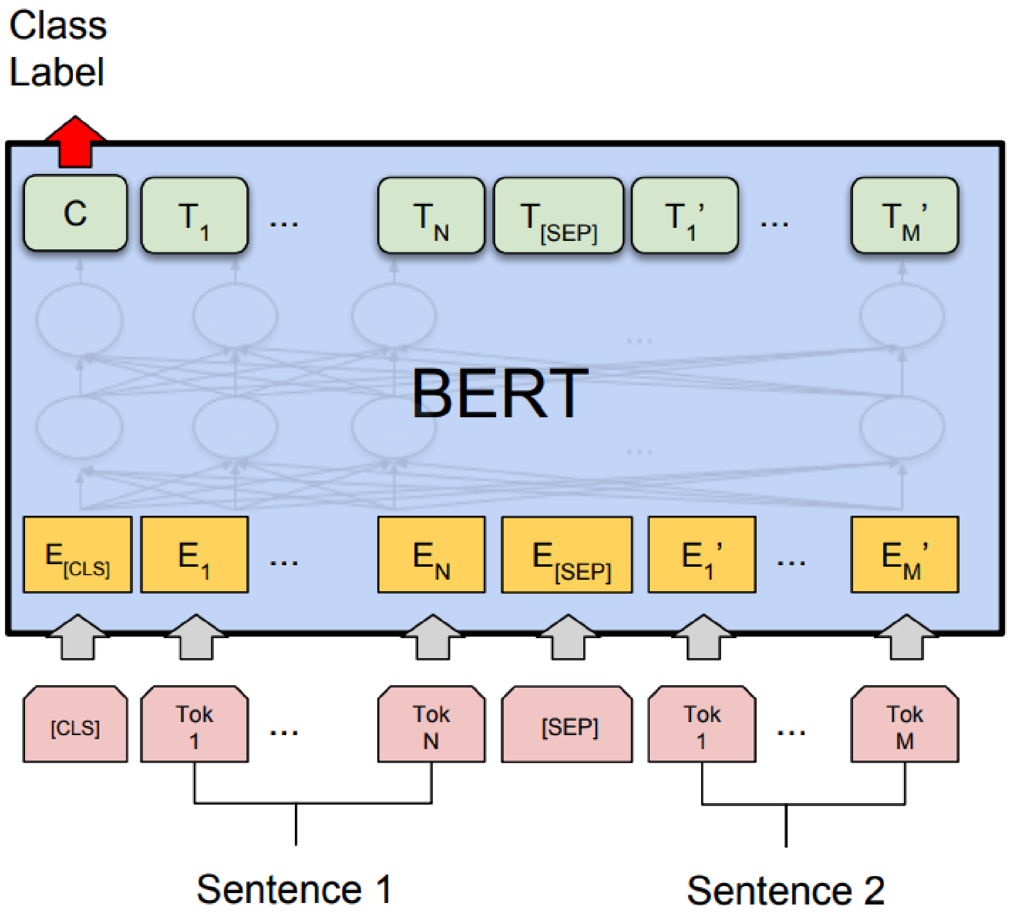
Natural Language Inference (NLI)

NLI is the task of **determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral)** given a "premise". [1]

Natural Language Inference (NLI)

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

Natural Language Inference (NLI)



Natural Language Inference (NLI)

The loss is simply the cross entropy or the divergence over the output of the `CLS` token and the true label.

$$\mathcal{L}_{NLI} = \mathcal{L}_{CE_{CLS}}$$

We are trying to compress the information about both sentence in one `CLS` token via attention and decide about their relationship.

Is it possible to help the model infering more information with les text data?

Going Further: LM as Knowledge Graphs

Dragon

Questions?

References

[1] https://paperswithcode.com/task/natural-language-inference

[2] Singla, S., & Feizi, S. (2021). Causal imagenet: How to discover spurious features in deep learning. arXiv preprint arXiv:2110.04301, 23.

[3] Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., & Liang, P. S. (2019). Unlabeled data improves adversarial robustness. Advances in neural information processing systems, 32.

[4] Pretrained Transformers Improve Out-of-Distribution Robustness (Hendrycks et al., ACL 2020)

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[6] Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning (pp. 12697-12706). PMLR.