CamemBERT: a Tasty French Language Model

EDF 08/09/2020

Presented by Benjamin Muller

Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*,
Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah and Benoît Sagot.

NLP at INRIA Paris - ALMAnaCH team

Who am I ?

- Researcher (PhD student) at INRIA in the AlmanaCH project team

What do I do ?

- Focus on **transfer learning** for Out-Of-Distribution scenarios (non canonical data, cross-lingual transfer...)

The ALMAnaCH team

- We build **linguistics ressources, data sets**, parsing models**, language models** and release **everything to the public:** Oriented toward **modelling language variability**
- We work with many startups, universities and companies (from Hyperlex to Facebook, from Bar Ilan University to Stanford, part of the PRAIRIE institute)

Outline

I. **Background: NLP & Language Models**

II. What is CamemBERT ? (quick demo)

III. CamemBERT on Downstream Tasks

IV. CamemBERT in Practice
   - Domain Adaptation
   - In production

V. NLP Beyond Language Models

What’s Natural Language Processing?

NLP aims at structuring language productions

- in **minimal sense unit** : words, morphemes..
- in **syntactic unit/relation** : subject, verb, object, modifier
- in **semantic unit**: who did what to whom? who did say what?

This structuring implies the definition of these units as well as their scopes

- “word” vs token: **chépa, ’la pas [cassé sa pipe] lui deja, wsh**⇒ Typographic segmentation doesn’t hold
- regular vs non-canonical syntax: ***John is tired*** vs ***dunno too tired 2think***
  ⇒ Who is tired? the speaker or someone else?
- The context of a production: **I don’t feel that brand and stuff.**
  ⇒ What brand? what stuff? who is he answering to?

How does it work?

Using linguistics knowledge. One principle, two schools:

- **Building rules (grammars)** and associated software.
  ⇒ Old-school approach, costly. Precise but very **application-dependant.**
- **Building annotated data set and supervised models will do the same** as (^) but better (need a labelled dataset per task x domain → 1 model per task x domain)
  ⇒ Data-driven approach, we try to generalise the data. Flexible but domain sensitive

No (or much fewer) linguistics knowledge.

(i) Building « nothing » and counting on massive amount of data
to detect regularities, bring out information

➔ **Unsupervised approach** (=no prior linguistics knowledge)

(ii) Using (i) via language models and directly transfer knowledge to specific tasks

➔ **This is the current NLP revolution**

Representing text into vectors

Goal: How to build useful model representation that capture words meaning ?

Example: What is the meaning of " bardiwac"?

He handed her a glass of bardiwac.
Beef dishes are made to complement the bardiwacs.
Nigel staggered to his feet, face flushed from too much bardiwac.
Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
I dined off bread and cheese and this excellent bardiwac

Distributional Hypothesis: "Words in similar context tend to have similar meanings"
Haris, 1954

➔ **idea:** *Model context to Model words*

Brief History of representation models

- **Word Embeddings:**
  - Word2vec/Glove models build a static vectorial representation of words
  - Fits very well with **task-specific deep learning architecture** (great precision)
  - **Problems:** What about polysemy ? What to do with a new word ?
- **Solution**: **Contextualized Word Embeddings.**
  - **Idea:** Use a **neural language model** to provide a **context-dependent representation**.
  - Many models have appeared: Ulmfit, Elmo, **BERT**...
- Replace task-specific architectures with **Transfer Learning.**
  - **Fine-tuned directly** on downstream tasks.
  - Achieves **state-of-the-art performance** in many tasks.

Outline

I. Background: NLP & Language Models

II. What is CamemBERT ? (quick demo)

III. CamemBERT on Downstream Tasks

IV. CamemBERT in Practice
   - Domain Adaptation
   - In production

V. NLP Beyond Language Models

Most Ressources and Experiments in English Only

- **BERT** and variants very impactful but **mostly for English**
- What about **other languages**?
    - **Only multilingual models** such as mBERT, XLM and XLM-R but mBERT still lagging behind monolingual counterparts.
- Do **BERT performance boosts transfer** to **other languages**?

➔ Let’s find out on **French** with **CamemBERT**!

What is CamemBERT ?

CamemBERT is a **Transformer-based architecture** trained as a **Mask-Language Model** on **130GB** of French **OSCAR** (web) data

Ref: CamemBERT is based on BERT (Devlin et. al 2018) and Roberta (Liu et. al. 2019)

CamemBERT: Objective Function

- **15%** of tokens are MASKED
- The model **learns to predict** <mask> tokens based on the (bidirectional) context

<s> Le CamemBERT est un <mask> délicieux </s>

Encoding

france

femme

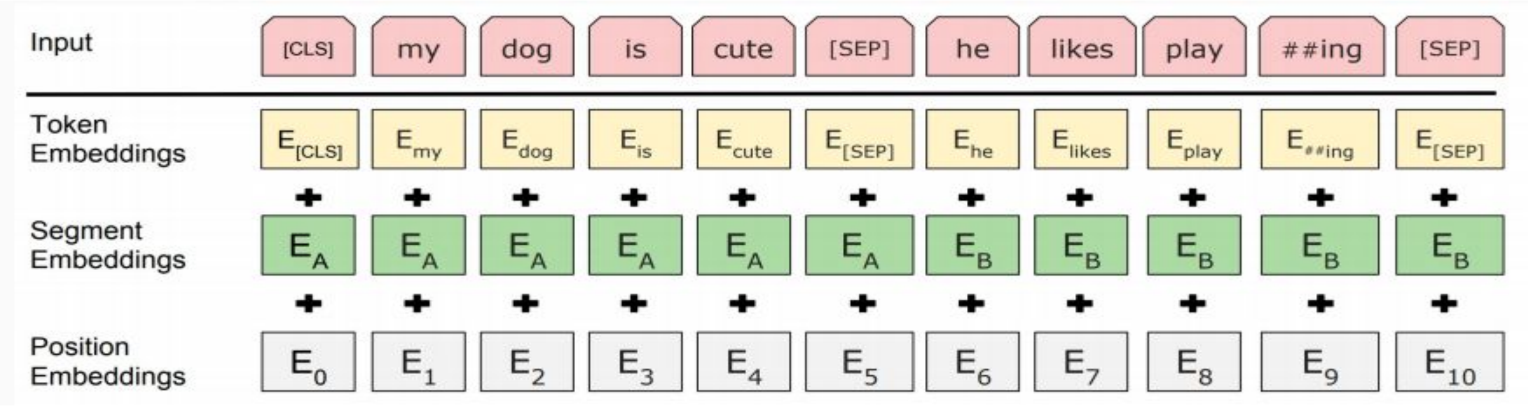
fromage

bon

....

Learn to Predict most likely token based on the observations

CamemBERT: Input Representation



(Devlin et. al 2018)

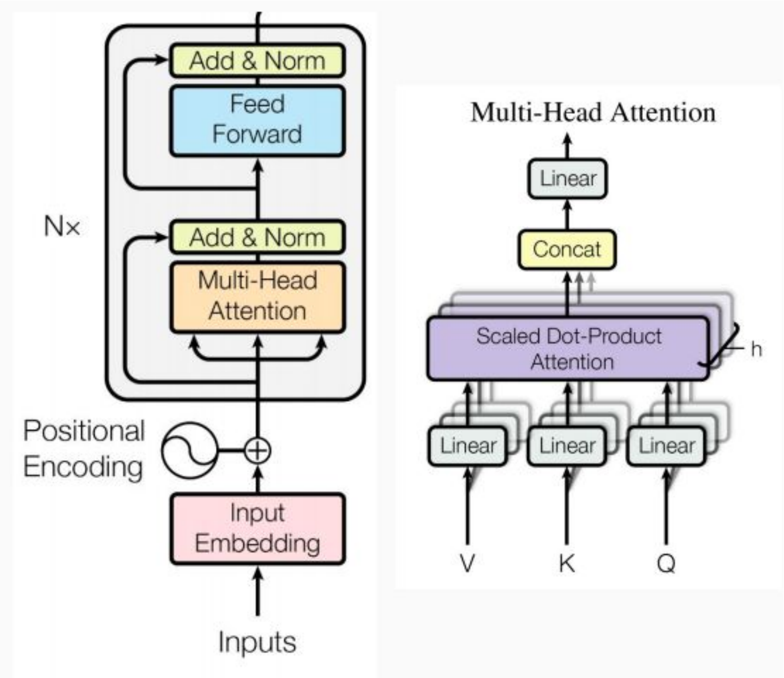
- Text is split with **Sentencepiece tokenization** (Kudo et. al. 2018):
  Unlikely words are split into sub-words unit ⇒ No Out-of-Vocabulary words
- CamemBERT uses a 32k tokens vocabulary
- Each token is input as **the sum of three embedding vectors** (position, token, segment)

CamemBERT: Architecture

- A Transformer is a stack of **self-attention layers** followed by **dense** layers

- CamemBERT-base is 12 layers
  CamemBERT-Large is 24 layers

- Non-recurrence operations makes it very computationally efficient
  (for GPUs)



(Vaswani et al. 2017)

CamemBERT: Trained on Open French Data

CamemBERT is trained on OSCAR.

- **OSCAR** is a clean **extract** of **Common Crawl** (Ortiz et al., 2019).
- **Open-source and freely available** at oscar-corpus.com.
- **French** data: **138GB of text**, 32.7B tokens, 59.4M documents.
- **Heterogeneous** data with **diverse styles** and **domains**.

➔ Very **COSTLY** to train (lots of GPU-hour)

(demo)

Outline

I. Background: NLP & Language Models

II. What is CamemBERT ? (quick demo)

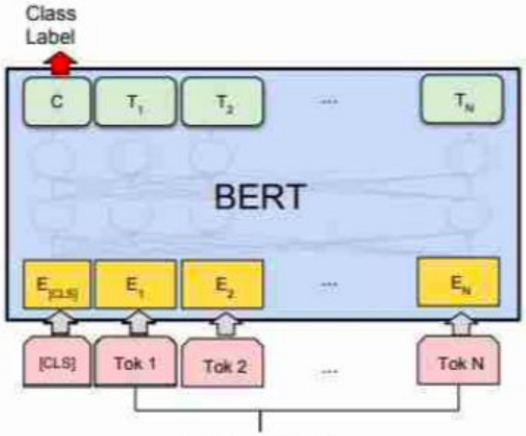
III. CamemBERT on Downstream Tasks

IV. CamemBERT in Practice
- Domain Adaptation
- In production

V. NLP Beyond Language Models

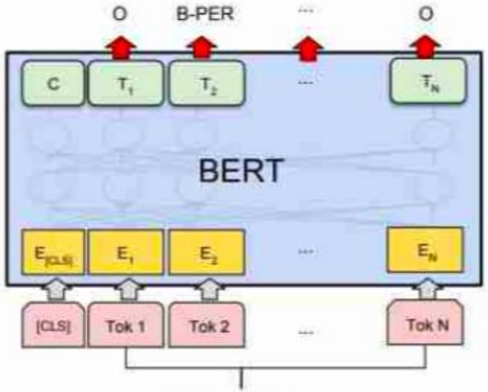
CamemBERT fine-tuning for downstream tasks

Sequence Classification



(b) Single Sentence Classification Tasks: SST-2, CoLA

Sequence Labelling



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Evaluation

Tasks and baselines:

- **Part-Of-Speech Tagging (POS)**: mBERT, XLM, UDify, and UDPipe Future
- **Dependency Parsing**: Same as POS tagging
- **Named Entity Recognition (NER)**: CRF, BiLSTM-CRF, and mBERT.
- **Natural Language Inference**: mBERT, XLM, and XLM-R

Two evaluation settings:

- **Fine-tuned**: CamemBERT is fine-tuned on the downstream tasks
- **As Embeddings**: Freeze CamemBERT, use output embeddings as input to another model

Tagging and Parsing: SotA Performance

- **Word labelling** and **structure prediction** tasks
- Evaluation on **4** Universal Dependencies **treebanks** of **different genres.**
- Fine-tuned models get **state-of-the-art results** on almost all datasets.
    - *Spoken has no punctuation, no uppercasing, much more difficult to*

Model	GSD		Sequoia		Spoken		ParTUT	
	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS
mBERT (fine-tuned)	97.48	89.73	98.41	91.24	96.02	78.63	97.35	91.37
$XLM_{MLM\text{-}TLM}$ (fine-tuned)	98.13	90.03	98.51	91.62	96.18	80.89	97.39	89.43
UDify (Kondratyuk, 2019)	97.83	91.45	97.89	90.05	96.23	80.01	96.12	88.06
UDPipe Future (Straka, 2018)	97.63	88.06	98.79	90.73	95.91	77.53	96.93	89.63
+ mBERT + Flair (emb.) (Straka et al., 2019)	97.98	90.31	**99.32**	93.81	**97.23**	81.40	97.64	92.47
CamemBERT (fine-tuned)	**98.18**	**92.57**	99.29	**94.20**	96.99	81.37	**97.65**	**93.43**
UDPipe Future + CamemBERT (embeddings)	97.96	90.57	99.25	93.89	97.09	**81.81**	97.50	92.32

NER, NLI & QA: Drastic Improvements

- Word labelling (NER), sequence classification (NLI) and QA (FQuAD)
- **State-of-the-art results** on **all tasks.**

Model	FQuAD1.1-test F1	FQuAD1.1-test EM	FQuAD1.1-dev F1	FQuAD1.1-dev EM
Human Perf.	91.2	75.9	92.1	78.3
CamemBERT$_{\text{BASE}}$	88.4	78.4	88.1	78.1
CamemBERT$_{\text{LARGE}}$	**92.2**	**82.1**	**91.8**	**82.4**
FlauBERT$_{\text{BASE}}$	77.6	66.5	76.3	65.5
FlauBERT$_{\text{LARGE}}$	80.5	69.0	79.7	69.3
mBERT	86.0	75.4	86.2	75.5
XLM-R$_{\text{BASE}}$	85.9	75.3	85.5	74.9
XLM-R$_{\text{LARGE}}$	89.5	79.0	89.1	78.9

Table 9: Results of the experiments for various monolingual and multilingual models carried out on the training dataset of **FQuAD1.1-train** and evaluated on test and development sets of FQuAD1.1

Model	Acc.	#Params
mBERT (Devlin et al., 2019)	76.9	175M
XLM$_{\text{MLM-TLM}}$ (Lample and Conneau, 2019)	80.2	250M
XLM-R$_{\text{BASE}}$ (Conneau et al., 2019)	80.1	270M
CamemBERT (fine-tuned)	**82.5**	110M
Supplement: LARGE models		
XLM-R$_{\text{LARGE}}$ (Conneau et al., 2019)	85.2	550M
CamemBERT$_{\text{LARGE}}$ (fine-tuned)	**85.7**	335M

NLI

Model	F1
SEM (CRF) (Dupont, 2017)	85.02
LSTM-CRF (Dupont, 2017)	85.57
mBERT (fine-tuned)	87.35
CamemBERT (fine-tuned)	89.08
LSTM+CRF+CamemBERT (embeddings)	**89.55**

NER

Crucial questions: How Much Training Data?

- **4GB vs. 138GB**

⇒ **Competitive results with** as few as **4GB of data** for a Base model!

Proves that strong models can be trained even on low resource languages or domain-specific datasets.

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		Average		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55

CamemBERT, a Useful Resource

- CamemBERT paved the way for other non-english monolingual models.
    - Since pre-publication, many other models have come out (FlauBERT for French, BERTje for Dutch, FinBERT for Finnish, PhoBERT for Vietnamese...).

- CamemBERT **models** are **open-source** and ready to use.
    - **100k+** downloads since its released
    - Models used for processing French legal text (Bennesty et al. 2019; Chavallard et al. 2020), French financial data, French Question Answering (d' Hoffschmidt et al. 2020; Keraron et al. 2020)...

CamemBERT in practice: How to use it ?

Check out camembert-model.fr for more details!

Available in **HuggingFace** and **Fairseq**

- Easy to load/to fine-tune/for prediction

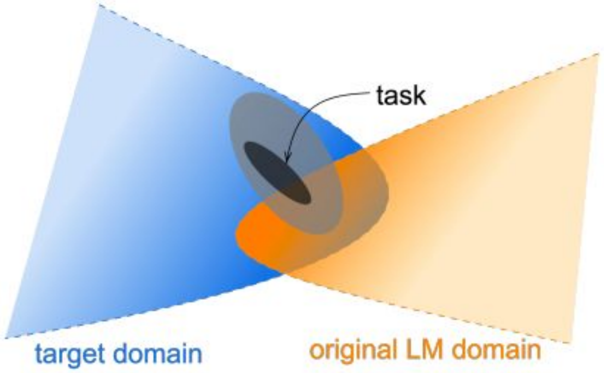
```
> import torch
> camembert =
torch.hub.load('pytorch/fairseq',
'camembert')
> camembert.eval()  # disable dropout
> masked_line = 'Le camembert est <mask> :)'
> camembert.fill_mask(masked_line, topk=3)

[('Le camembert est délicieux :)', 0.4909, '
délicieux'),
 ('Le camembert est excellent :)', 0.1056, '
excellent'),
 ('Le camembert est succulent :)', 0.0345, '
succulent')]
```

CamemBERT for your Domain

"Don't stop pretraining" (Gururangan et al., 2020)

- ➔ Language Models can be adapted by simply **fine-tuning** them in an unsupervised way using their Mask-Language Model objective **on a new domain**
- ➔ Improve performance of up to **+3 points** in downstream classification



CamemBERT in production: How to speed up CamemBERT?

Train a smaller version of BERT using **Knowledge Distillation** (Hinton 2015, Sanh 2020)

Distillation consists in training a ***student model*** based on the prediction of a ***teacher model***

$$L = -\sum_i t_i * log(s_i)$$

With **t** the logits from the teacher and **s** the logits of the student

- ➔ Train a smaller Transformer using CamemBERT as a ***teacher*** (up to x5 speed up)
- ➔ No Distil-Camembert available <u>for now</u> !

Take Home Message

- **CamemBERT** achieves **state-of-the-art** results in **5 downstream tasks**
  - Surpass multilingual models and confirms what was found for English monolingual language models
- **Type of data matters:** Pretraining on **heterogeneous data** is **important**
  - Common Crawl better than Wikipedia
- **Size doesn't matter much: Strong models** can be trained with as **little as 4GB of raw text**
  - Good news for low-resource languages or domain-specific data.
- **Speed-up** possible with more compact distilled models

Beyond Language Modeling

- Current language-model based **pretraining-fine-tuning** is impressive (a single architecture for all tasks)
- A few **very large language models** trained and shared freely **adapted** and **analysed** by the research community and by companies (for how long?)
- Very poor understanding of how and what these models capture: Research in Bertology

What comes next ? scaling even more ?

- Training Ever-Larger Models (Monolingual and Multilingual)
- New framework: Language Models as few shot learners: GPT-3 (x1000 more parameters than Camembert)

But can we learn it all from forms ?

- **"Language Models do not learn meaning"** (Bender et. al., 2020)

➔ Multi-Modal approaches (ERNIE Zhang et al. 2019, Hill et. al 2020)

Thank you!