école
normale
supérieure
paris—saclay

SNLP - Project presentation

CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation

Marine Astruc, Josselin Dubois, Javier Ramos-Gutiérrez, Gabriel Watkinson

Paris-Saclay University

October 17, 2023

Summary

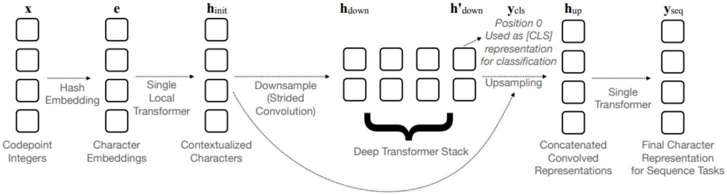
1. Introduction

2. Reproduction of the paper
   - Information Seeking
   - Named Entity Recognition

3. Further experiments
   - Entailment Analysis
   - Translation

4. Conclusion

Introduction



x
e
h_{init}
h_{down}
h'_{down}
y_{cls}
h_{up}
y_{seq}
Position 0 Used as [CLS] representation for classification
Hash Embedding
Single Local Transformer
Downsample (Strided Convolution)
Upsampling
Single Transformer
Deep Transformer Stack
Codepoint Integers
Character Embeddings
Contextualized Characters
Concatenated Convolved Representations
Final Character Representation for Sequence Tasks

Figure: CANINE Neural Architecture

- Character-level model (tokenizer-free)
- Efficient downsampling

Two pre-trained models

- CANINE-C: Auto-regressive Character loss
- CANINE-S: Subword loss

Information Seeking

- Datasets: TyDi QA (primary task) dataset of information-seeking questions in 11 typologically diverse languages (Japanese, Arabic, English, ...)
- Models : CANINE-C and CANINE-S
- Evaluation : SQUAD metric

	CANINE C		**CANINE S**	
	Ours	Paper	Ours	Paper
Exact matches	19%	N/A	21%	N/A
F1-Score	**72.9**	65.7	**76.1**	66.0

Named Entity Recognition

- Datasets:
  - CoNLL. 3 European languages: Spanish, Dutch (2002), English (2003).
  - MasakhaNER. 10 African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof, Yorùbá.

- Models:
  - mBert
  - Canine-C
- Classes:
  - 'PER', 'ORG', 'LOC', 'MISC' (CoNLL only), 'DATE' (MasakhaNER only).
  - 'B-' (beginning), 'I-' (intermediate).

[	Cornellá	de	Llobregat	(	Barcelona	)	,	23	may	(	EFE	)	.	]
[	B-LOC	I-LOC	I-LOC	O	B-LOC	O	O	O	O	O	B-ORG	O	O	]

Named Entity Recognition

CoNLL	Paper		Ours	
Language	mBERT	CANINE-C	mBERT	CANINE-C
Dutch	90.2	74.7	90.3	87.0
English	91.1	79.8	90.3	89.9
German	82.5	64.1	-	-
Spanish	87.6	77.4	87.1	88.6
Macro Avg	87.8	74.0	89.3*	88.5*

Table: F1 score on CoNLL test sets.

Named Entity Recognition

MasakhaNER	Paper		Ours	
Language	mBERT	CANINE-C	mBERT	CANINE-C
Amharic	0.0	44.6	0.0	15.6
Hausa	89.3	76.1	78.2	69.7
Igbo	84.6	75.6	76.5	69.6
Kinyarwanda	73.9	58.3	61.7	45.3
Luganda	80.2	69.4	64.7	59.9
Luo	75.8	63.4	27.6	15.7
Nigerian Pidgin	89.8	66.6	82.7	71.1
Swahili	87.1	72.7	83.0	68.2
Wolof	64.9	60.7	57.8	54.6
Yorùbá	78.7	67.9	69.3	52.4
Macro Avg	72.4	65.5	60.2	52.2

Table: F1 score on MasakhaNER test sets.

Entailment analysis

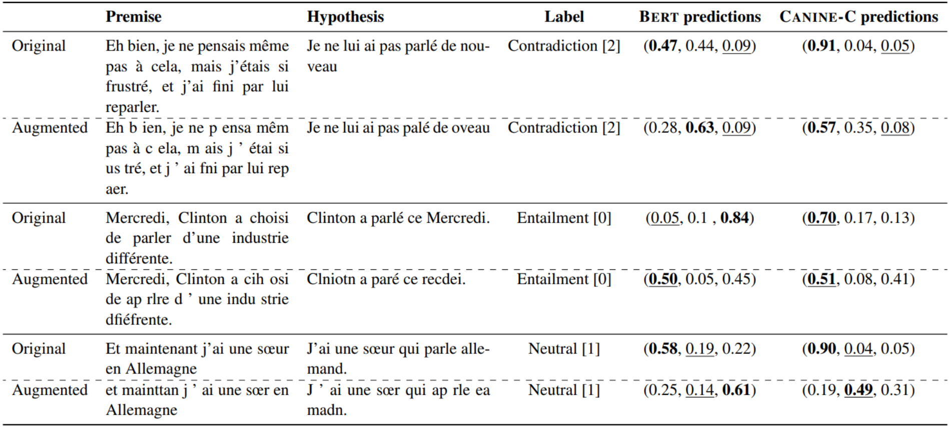
- **Dataset:** XNLI
  - Translation of MNLI, in 14 languages, 400k pairs
  - NLI : predict if two sentences are in agreement, disagreement or neutral
- **Model:** 3 label classification
  - Baseline: pretrained multilingual BERT
  - Evaluation: pretrained CANINE
  - CANINE-C on character level loss
  - CANINE-S on subword level loss

Entailment Analysis

Model	Train languages	English	Bulgarian	German	Greek
BERT	**0.673**	**0.707**	0.653	**0.617**	**0.597**
Canine-C	0.667	0.703	**0.667**	0.475	0.474
Canine-S	0.654	0.676	0.658	0.458	0.447

Table: F1 score on different test sets. The F1 score is a weighted average between the 3 class and languages for the first column, and is evaluated on a test set of 5k observations never seen during training. The train languages are English, French, Spanish, Bulgarian and Russian.

Entailment Analysis



	Premise	Hypothesis	Label	BERT predictions	CANINE-C predictions
Original	Eh bien, je ne pensais même pas à cela, mais j'étais si frustré, et j'ai fini par lui reparler.	Je ne lui ai pas parlé de nouveau	Contradiction [2]	(**0.47**, 0.44, 0.09)	(**0.91**, 0.04, 0.05)
Augmented	Eh b ien, je ne p ensa mêm pas à c ela, m ais j ' étai si us tré, et j ' ai fni par lui rep aer.	Je ne lui ai pas palé de oveau	Contradiction [2]	(0.28, **0.63**, 0.09)	(**0.57**, 0.35, 0.08)
Original	Mercredi, Clinton a choisi de parler d'une industrie différente.	Clinton a parlé ce Mercredi.	Entailment [0]	(0.05, 0.1 , **0.84**)	(**0.70**, 0.17, 0.13)
Augmented	Mercredi, Clinton a cih osi de ap rlre d ' une indu strie dfiéfrente.	Clniotn a paré ce recdei.	Entailment [0]	(**0.50**, 0.05, 0.45)	(**0.51**, 0.08, 0.41)
Original	Et maintenant j'ai une sœur en Allemagne	J'ai une sœur qui parle allemand.	Neutral [1]	(**0.58**, 0.19, 0.22)	(**0.90**, 0.04, 0.05)
Augmented	et mainttan j ' ai une sœr en Allemagne	J ' ai une sœr qui ap rle ea madn.	Neutral [1]	(0.25, 0.14, **0.61**)	(0.19, **0.49**, 0.31)

Figure: Case study of the NLI task with 3 random examples.

Entailment Analysis

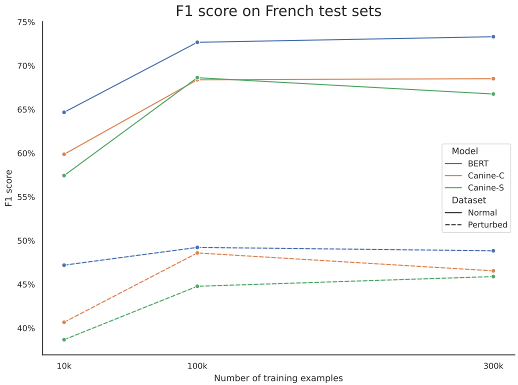


Figure: Performance of the three French models trained on 10k, 100k and 300k observations, evaluated using weighted F1 on the original and augmented French test set.

Translation

- Dataset: WMT14 (French, Czech, Hindi, ...)
- Model: Encoder-Decoder
    - Encoder: CANINE (frozen)
    - Decoder: Bart (fine-tuned)

- Problems:
    - Embedding too large to be computed (1,114,112 unicode characters) $\longrightarrow$ restrict on latin-characters,
    - Some words are not correct.
- Success:
    - Generates $\pm$ correct sentences
    - Perceives sentences structures (dialogue, ...)

Translation

Input	Ground truth	Canine output
”You saw?” he said.	–Tu as vu? dit-il.	– Vous êtes prises, dit Arthos Conseigne a Marguteries.
”We shall have to beat the forest,” said the engineer, ”and rid the island of these wretches.	– Il faudra battre la forêt, dit l’ingénieur, et débarrasser l’île de ces misérables.	Ils étaient de la maison, les compagniers, et se regardait avec une chambre explication dans longtemps du souvert l’autres.
Phileas Fogg, having shut the door of his house at half-past eleven, and having put his right foot before his left five hundre	Phileas Fogg avait quitté sa maison de Saville-row à onze heures et demie, et, après avoir placé cinq cent soixante-quinze foi	Il avait été par une fois, et les conseillement de cette heure, il faut pour la plusie dans sous longtemps du bonher se rappel

Conclusion

- Interesting/Important to work on characters
- Good results on a wide variety of tasks
- Still limitations (sequence generation, ...)

- Might be interesting to compare to character-levels tokenizer (sentencepiece, ...)