Cross-Lingual Learning vs. Low-Resource Fine-Tuning: A Case Study with Fact-Checking in Turkish

Recep Firat Cekinel, Pinar Karagoz, Cagri Coltekin

Motivation

- Fact-checking aims to assess the truthfulness of statements

- Fake stories disseminate faster than true information on social media (Vosoughi et al., 2018) to manipulate public opinion on major events (i.e. Brexit referendum (Pogue, 2017), 2016 US Presidential Elections (Allcott and Gentzkow, 2017)

Hire factcheckers to fight election fake news, EU tells tech firms

Parliamentary elections thought vulnerable to fake news will test social media firms and bloc's new DSA laws

Recent elections in EU states have been subject to online disinformation. Photograph: Jean-François Badias/AP

from Getty Images

Contribution

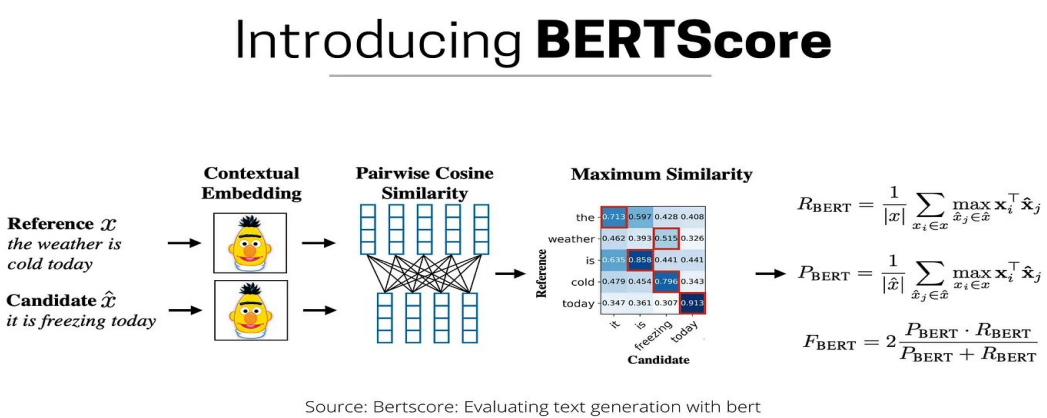
- Assessing the effectiveness of **cross-lingual transfer learning** for low-resource languages, with a particular focus on Turkish

- Releasing a **Turkish fact-checking dataset** obtained by crawling three Turkish fact-checking websites

- Presenting experimental results, comparing **zero- and few-sho**t prompt learning and **fine-tuning** on large language models and underscoring the need to utilize a small amount of native data

FCTR Dataset

- Turkish fact-checking dataset
- Teyit, Dogrulukpayi and Dogrula listed on the Duke Reporters' Lab and members of the International Fact-Checking Network (IFCN)
- Claim, evidence, summary, label, URL, publication date
- BERTScore (Zhang et al., 2019b) to identify candidate duplicate claims



Source: Bertscore: Evaluating text generation with bert

Code for Bertscore is available at https://github.com/Tiiiger/bert_score

FCTR Statistics

- 3238 claims dating from 23.07.2016 to 11.07.2023
- Dogrulukpayi: 742 claims
- Dogrula : 525 claims
- Teyit: 1971 claims

Veractiy Labels	Sources	Counts
false	Dogrula, Teyit, Dogrulukpayi	2780
true	Dogrula, Teyit, Dogrulukpayi	203
mixed	Teyit	109
partially false	Dogrulukpayi	72
unproven	Teyit	37
half true	Dogrula	17
mostly false	Dogrula	14
mostly true	Dogrula	6

Table 1: Number of veracity labels in the FCTR dataset

Snopes Dataset

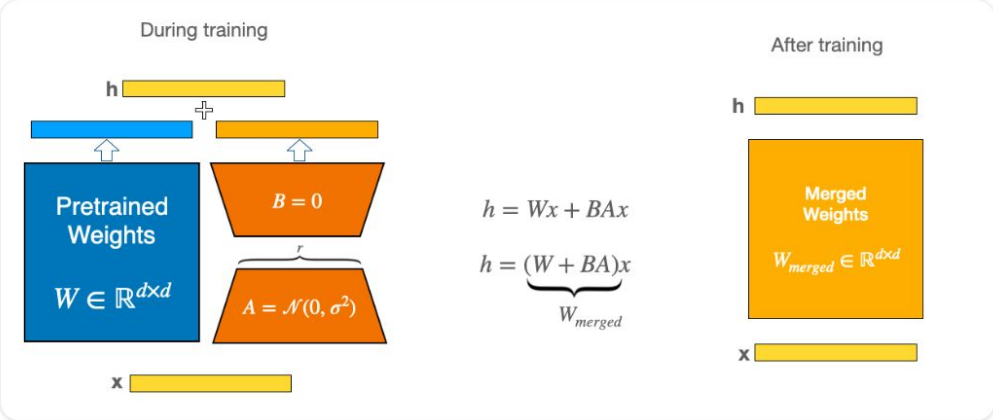
- Independent organization committed to fact-checking in English
- Covers a broad range of topics, including politics, health, science, popular culture, etc.
- 6402 claims retrieved from November 24, 1996 to August 17, 2023

Veractiy Labels	Counts
false	2270
true	1467
mixture	588
miscaptioned	375
unproven	284
labeled satire	283
correct attribution	247
mostly false	237
mostly true	198
other	453

Table 2: Number of veracity labels in the Snopes dataset[9]

Model

- Llama 2 (Touvron et al., 2023)[1]
  - Three variants, with parameter sizes of 7 billion, 13 billion, and 70 billion
  - The pre-training data includes information up to September 2022, while the fine-tuning data is up to June 2023
- Parameter efficient fine-tuning with QLoRA (Dettmers et al., 2023)



1 https://huggingface.co/meta-llama

Instruction Prompting

- Instruction tuning significantly improves the performance of large language models across a range of tasks (Zhang et al., 2023)
- Prompting was shown to be an effective way to describe models' reasoning steps by enabling the generation of coherent reasoning chains leading to the desired output (Wei et al., 2022)
- Alpaca prompt template (Taori et al., 2023)

```
### Instruction: Is the following statement ”true” or ”false”?
### Input:
A series of photographs show the skeletal remains of the biblical giant Goliath.
### Response:
false
```

Experimental Setup

- Datasets: Snopes and FCTR
  - FCTR500 (203 True - 297 False) and FCTR1000 (203 True - 797 False)
  - 80% of the data for training, 10% for validation, and 10% for testing

- Baselines
  - **SVM** (Cortes and Vapnik, 1995): sparse word and n-gram features weighted by tf-idf.
  - **multilingual BERT (mBERT)** (Devlin et al., 2019): weighting cross-entropy loss with the inverse class ratios, causing the model to assign a larger penalty to the majority class

- PeFT: QLoRA
  - Low rank matrices (r) = 16
  - scaling factor for the weight matrices (lora_alpha) = 64
  - Lora_dropout = 0.1

Snopes Fine Tuning Results

Input	Model	F1-macro	F1-binary
claim 10-fold	SVM	0.651	0.709
claim	SVM	0.695	0.763
claim	mBERT	0.705	0.802
claim	LLaMA-7B	0.766	0.838
claim	LLaMA-13B	0.814	0.866
claim	LLaMA-70B	**0.826**	**0.890**

Table 3: Veracity prediction on the Snopes data

FCTR Fine Tuning Results

Input	Model	F1-macro	F1-binary
claim 10-fold	SVM-fctr500	0.682	0.610
claim	SVM-fctr500	**0.714**	0.709
claim	mBERT-fctr500	0.653	0.750
claim	Llama-7B-fctr500	0.632	0.765
claim	Llama-13B-fctr500	0.635	0.679
claim	Llama-70B-fctr500	0.649	**0.783**
+summary	mBERT-fctr500	0.752	0.861
+summary	Llama-13B-fctr500	**0.890**	**0.923**

Table 4: Fine tuning on the FCTR500 data

Input	Model	F1-macro	F1-binary
claim	SVM-fctr1000	**0.671**	0.842
claim	mBERT-fctr1000	0.518	0.797
claim	Llama-7B-fctr1000	0.561	**0.864**
claim	Llama-13B-fctr1000	0.642	0.839
+summary	mBERT-fctr1000	0.729	0.902
+summary	Llama-13B-fctr1000	**0.828**	**0.947**

Table 5: Fine tuning on the FCTR1000 data

In-Context Learning Results

Input	Model	F1-macro	F1-binary
zero shot	mBERT	0.550	0.667
zero shot	Llama-7B	0.488 ∓ 0.026	0.577 ∓ 0.027
1-shot	Llama-7B	0.536 ∓ 0.006	0.742 ∓ 0.009
2-shot	Llama-7B	0.545 ∓ 0.035	0.632 ∓ 0.045
3-shot	Llama-7B	0.577 ∓ 0.011	0.642 ∓ 0.029
4-shot	Llama-7B	0.538 ∓ 0.021	0.609 ∓ 0.024
5-shot	Llama-7B	0.533 ∓ 0.021	0.647 ∓ 0.022
zero shot	Llama-13B	0.498 ∓ 0.014	0.699 ∓ 0.006
1-shot	Llama-13B	0.489 ∓ 0.026	0.683 ∓ 0.023
2-shot	Llama-13B	0.530 ∓ 0.028	0.689 ∓ 0.019
3-shot	Llama-13B	0.482 ∓ 0.022	0.670 ∓ 0.028
4-shot	Llama-13B	0.529 ∓ 0.036	0.638 ∓ 0.028
5-shot	Llama-13B	0.514 ∓ 0.013	0.632 ∓ 0.007
zero shot	Llama-70B	0.527 ∓ 0.042	**0.773** ∓ 0.016
1-shot	Llama-70B	0.507 ∓ 0.036	0.766 ∓ 0.018
2-shot	Llama-70B	0.539 ∓ 0.021	0.754 ∓ 0.013
3-shot	Llama-70B	0.492 ∓ 0.030	0.692 ∓ 0.023
4-shot	Llama-70B	0.542 ∓ 0.021	0.709 ∓ 0.014
5-shot	Llama-70B	**0.585** ∓ 0.017	0.709 ∓ 0.023

Table 7: Transfer learning on the FCTR500 data

Input	Model	F1-macro	F1-binary
zero shot	mBERT	0.529	0.736
zero shot	Llama-7B	0.479 ∓ 0.019	0.647 ∓ 0.018
1-shot	Llama-7B	0.501 ∓ 0.017	0.857 ∓ 0.013
2-shot	Llama-7B	0.518 ∓ 0.010	0.706 ∓ 0.006
3-shot	Llama-7B	0.501 ∓ 0.010	0.691 ∓ 0.024
4-shot	Llama-7B	0.512 ∓ 0.023	0.694 ∓ 0.024
5-shot	Llama-7B	0.502 ∓ 0.030	0.690 ∓ 0.048
zero shot	Llama-13B	0.502 ∓ 0.011	0.803 ∓ 0.006
1-shot	Llama-13B	0.550 ∓ 0.016	0.811 ∓ 0.014
2-shot	Llama-13B	0.539 ∓ 0.033	0.788 ∓ 0.020
3-shot	Llama-13B	0.533 ∓ 0.017	0.763 ∓ 0.016
4-shot	Llama-13B	0.537 ∓ 0.010	0.758 ∓ 0.010
5-shot	Llama-13B	0.533 ∓ 0.029	0.737 ∓ 0.021
zero shot	Llama-70B	0.521 ∓ 0.018	**0.865** ∓ 0.002
1-shot	Llama-70B	0.528 ∓ 0.011	0.858 ∓ 0.011
2-shot	Llama-70B	**0.560** ∓ 0.033	0.841 ∓ 0.012
3-shot	Llama-70B	0.536 ∓ 0.023	0.806 ∓ 0.018
4-shot	Llama-70B	0.520 ∓ 0.019	0.808 ∓ 0.016
5-shot	Llama-70B	0.521 ∓ 0.018	0.778 ∓ 0.015

Table 8: Transfer learning on the FCTR1000 data

Neural Machine Translation

Dataset	Model	F1-macro	F1-binary
fctr500	mBERT	0.561	**0.789**
fctr500	LLaMA-7B	**0.576** $\mp$ 0.014	0.782 $\mp$ 0.007
fctr500	LLaMA-13B	0.567 $\mp$ 0.018	0.739 $\mp$ 0.013
fctr500	LLaMA-70B	0.571 $\mp$ 0.015	0.771 $\mp$ 0.007
fctr1000	mBERT	0.485	0.840
fctr1000	LLaMA-7B	0.524 $\mp$ 0.011	0.847 $\mp$ 0.003
fctr1000	LLaMA-13B	0.573 $\mp$ 0.013	0.879 $\mp$ 0.004
fctr1000	LLaMA-70B	**0.581** $\mp$ 0.012	**0.883** $\mp$ 0.003

Table 9: Turkish to English machine translation results

Dataset	Model	F1-macro	F1-binary
fctr500	mBERT	0.532	**0.757**
fctr500	LLaMA-7B	0.523 $\mp$ 0.019	0.630 $\mp$ 0.023
fctr500	LLaMA-13B	0.544 $\mp$ 0.018	0.708 $\mp$ 0.006
fctr500	LLaMA-70B	**0.553** $\mp$ 0.025	0.725 $\mp$ 0.022
fctr1000	mBERT	0.474	0.826
fctr1000	LLaMA-7B	0.481 $\mp$ 0.023	0.705 $\mp$ 0.020
fctr1000	LLaMA-13B	0.552 $\mp$ 0.044	0.800 $\mp$ 0.024
fctr1000	LLaMA-70B	**0.556** $\mp$ 0.018	**0.832** $\mp$ 0.011

Table 10: English to Turkish machine translation results

Summary

- Small gains from the high-resource language in zero-shot and few-shot settings, where few-shot learning shows slight improvement over zero-shot
- Even a small amount of training data provides better results than zero- or few-shot approaches
- LLMs surpass the simple models (SVM and mBERT) on English with a large margin
- Success of small models (SVM) that rely only on surface cues on the FCTR data
- The comparatively smaller Turkish data during pretraining is possibly a factor in low scores of Llama when fine-tuned with Turkish

Thanks!

Any questions?

Recep Firat Cekinel

Email: rfcekinel@ceng.metu.edu.tr

Github: https://github.com/firatcekinel/FCTR