Can Synthetic Examples help Low Resource Document Classification?

Sam Shleifer sshleifer@gmail.com
(Looking for Research/Summer opportunities!)

Background

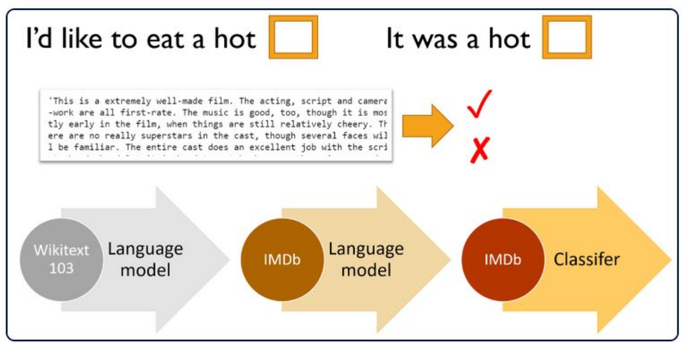
- Not many data augmentation for NLP papers: tough to implement during training + many ideas risk creating label noise.
- We augment IMDB movie reviews dataset with extra examples generated by two families of techniques:
  - Random token perturbations = messing with words in movie review
  - Backtranslation (BT) = translate whole review to a second language then back to English
- Experiments vary # synthetic and real examples model has access to
- Measure: did model correctly predict sentiment of movie review?

Augmented Examples

Operation	Sentence
None	A sad human comedy played out on the back roads of life.
BT (Spanish)	A sad human comedy that develops in the secondary roads of life.
BT (Bengali)	A sad man played the street behind comedy life.
Synonmym Replace†	A lamentable human comedy played out on the backward road of life.
Random Insert†	A sad human comedy played out on funniness the back roads of life.
Random Swap†	A sad human comedy played out on roads back the of life.
Random Delete†	A sad human out on the _ roads of life.

Table 1: BT stands for backtranslation. † Token Perturbation techniques from (Wei, 2019)

ULMFit



References:
- ULMFit: Howard and Ruder [2018]
- Backtranslation: Sennrich et al. [2016] and Edunov et al. [2018]
- Token Perturbations: Wei and Zou [2019]
- Virtual Adversarial loss: Miyato et al. [2016], Sato et al. [2018]

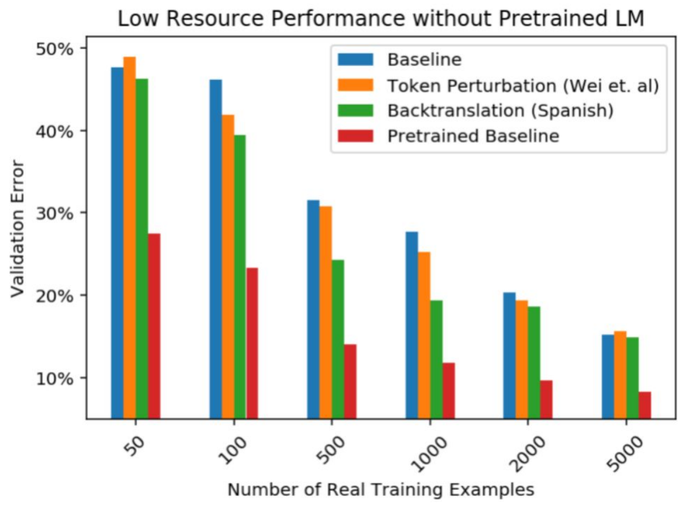
Results

- BT generates significant improvements when ULMFit only has access to a few examples, but stops working around 10,000 rows.. Token perturbations less effective.
- We train a model on only 50 real examples to 80.6% (+8%) by adding 500 synthetic backtranslated examples from 10 different languages!
- Using full IMDB Dataset (25K rows):
  - Training on synthetic examples doesn't help, but using them during inference time can contribute marginal (.13%) improvement.
  - Bigger ensemble can get up to 0.6% better than baseline model.

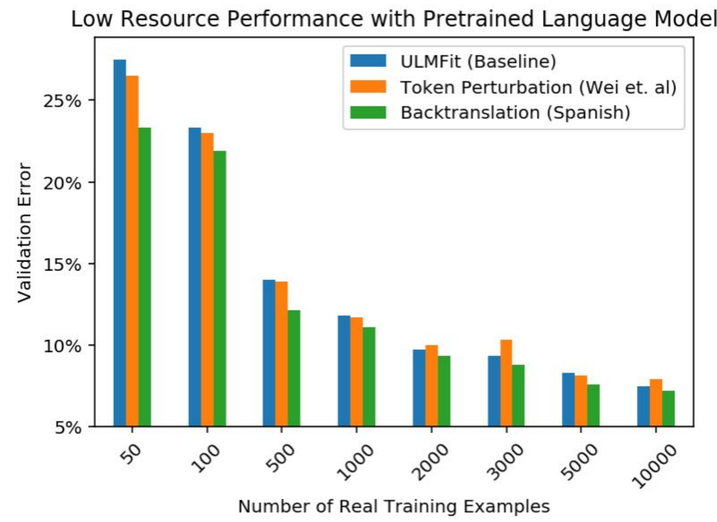
How much BT should I use?

Languages	Error@N=50	Error @N=1000
None	0.275	0.118
10 Languages	**0.194**	0.114
Spanish	0.233	0.111
Spanish, French	0.225	**0.109**
Spanish, French, Bengali	0.228	0.111
Bengali	0.241	0.113

Pretraining > Data Augmentation



Backtranslation Helps On Small Data



Full Dataset Results

- Training on synthetic examples stopped yielding improvements once the model had access to more than 15,000 examples
- Motivated testing whether using the BT examples as a form of test time augmentation might help the model.

Method	Reported Test Error	Replicated Test Error
(1) ULMFit FWD Howard and Ruder [2018]	5.30%	5.32%
(2) ULMFit BWD		7.38%
(3) (1) + (2)	**4.60%**	5.14%
(4) iVAT Sato et al. [2018]	5.66%	6.24%
ULMFit FWD + TTA *		5.10%
(3) + TTA*		4.97%
(3) + (4) + TTA *		**4.73%**

Table 3: * ensembles created for this project. The Replicated Test Error column is the test error when we run the authors published code without modification.

Can you guess the model's output?

A. `sentiment("6/10")`
B. `sentiment("6.5/10")`
C. `sentiment("7/10")`
D. `sentiment("Who ever came up with story is one sick person ... I'm only giving this movie a 9 because you FREAKED ME OUT FREAKS") high school")`