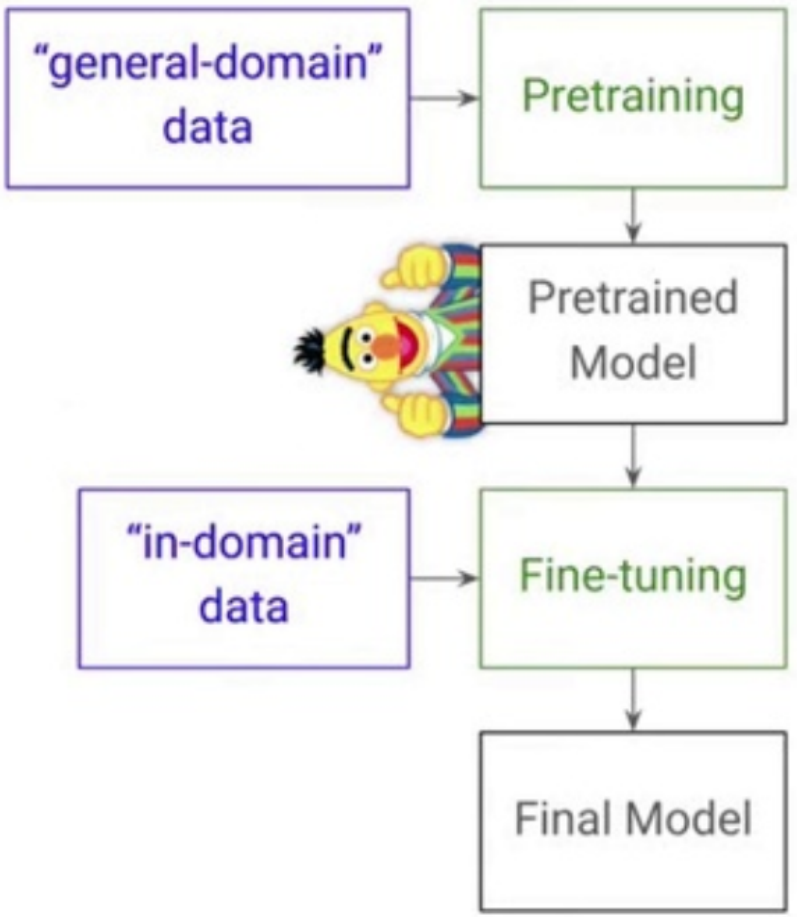
Unsupervised Domain Clusters in Pretrained Language Models

Roee Aharoni & Yoav Goldberg, ACL 2020

Abstract

- General training pipeline
- The notion of "in-domain" data is over-simplistic and vague
- Massive pre-trained LMs can cluster by domains without supervision $\rightarrow$ data-driven definition of domains
- Training on in-domain examples is better than training on all general examples
- Data selection methods: domain-cosine and domain-finetune

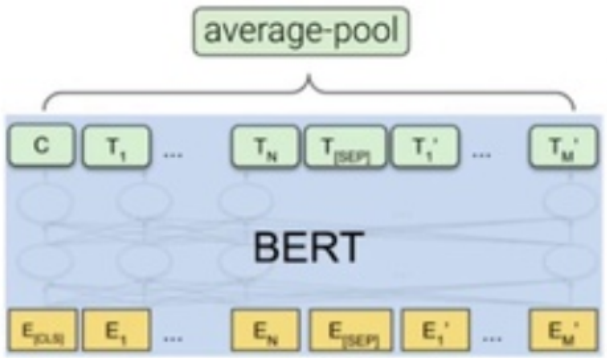


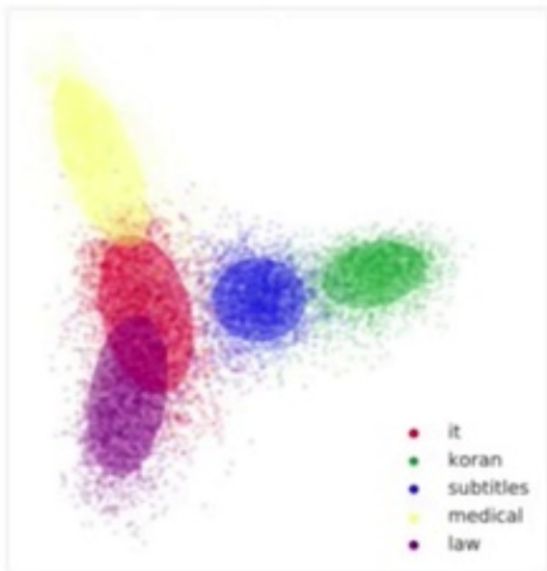
How do we define domains

- Data source? (TED, Paracrawl, books, ...)
- Topic (sports, finance, ...)
- Genre/Style? (spoken, scientific)
- All of the above?
- This work proposes data-driven approach to define domains

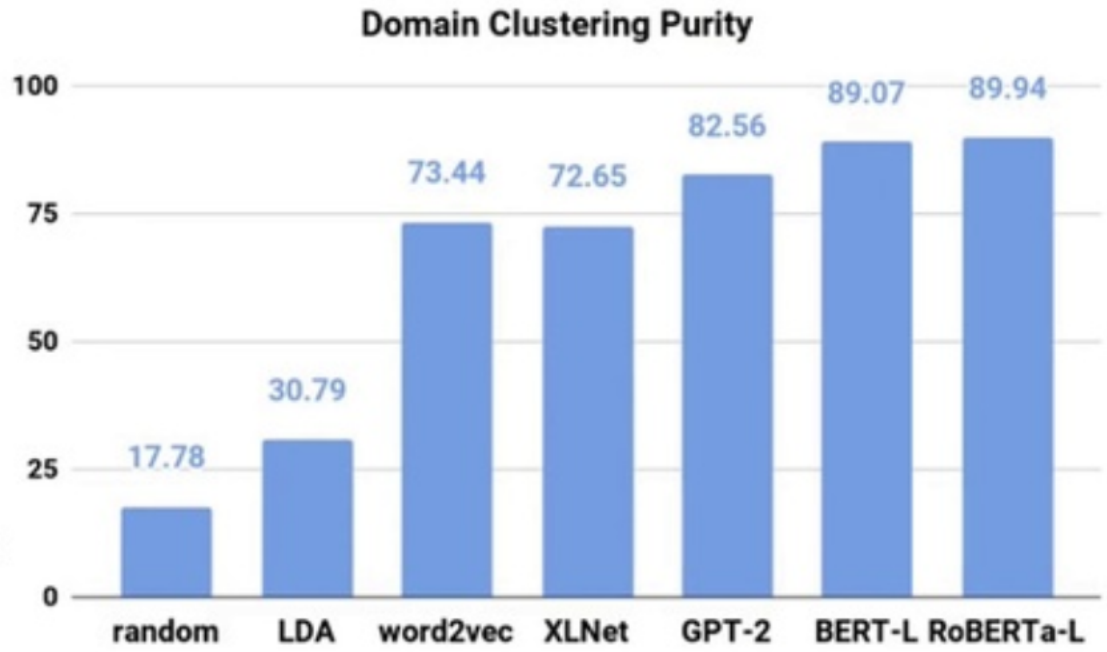
Unsupervised Domain Clusters in Pretrained LMs

- To test domain-cluster hypothesis, we sample 2000 English sentences from 5 domains: Medical, Legal, Koran, Movie Subtitiles, IT
- We then perform the following procedure with different pretrained models
  - **Encode** each sentence by final representation
  - **Cluster** the resulting 10k encoded sentences using Gaussian Mixture model
  - **Measure** the clustering purity vs. the source-based domain assignments



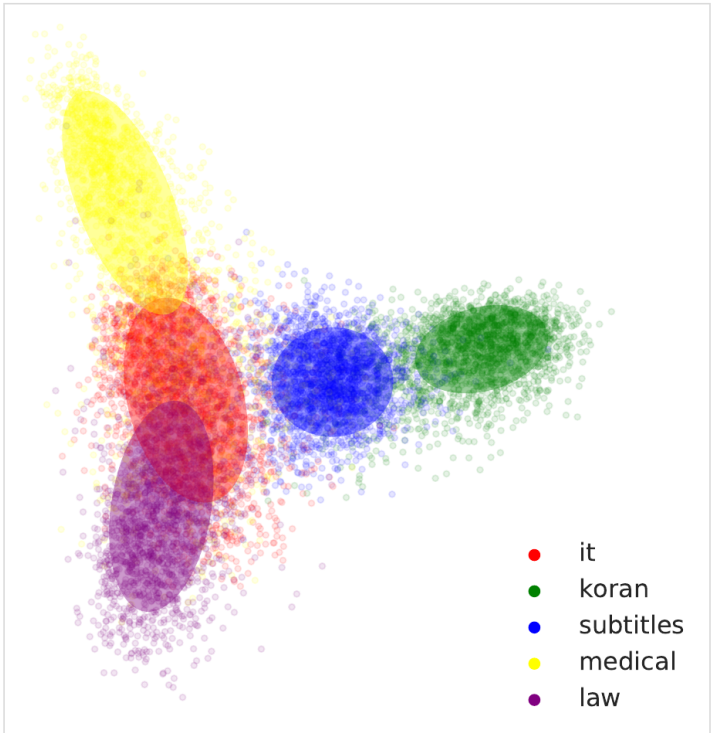


Unsupervised Clustering: Results



- Pretrained models are able to encode clusters, high purity scores
- Masked LMs are better at clustering than unmasked LMs

Unsupervised Clustering: Analysis



Subtitles assigned to IT
Push it up to the front of the screen.
Polyalloy requires programming to take permanent form.
Law assigned to Medical
- Viruses and virus-like organisms
where the glucose content is equal to or less than the fructose content.
Medical assigned to Law
This will be introduced by a Regulation adopted by the European Commission.
The marketing authorisation was renewed on 22 May 2002 and 22 May 2007.
IT assigned to Medical
R65: Harmful: may cause lung damage if swallowed
Automatic Red-Eye Removal

- There are some overlap between clusters
- Data-driven domain can be better than naive domain assignments based on the source, topics, genre or styles when collecting data
  - e.g., "...Viruses..." $\rightarrow$ Medical, "...Regularization..." $\rightarrow$ Law

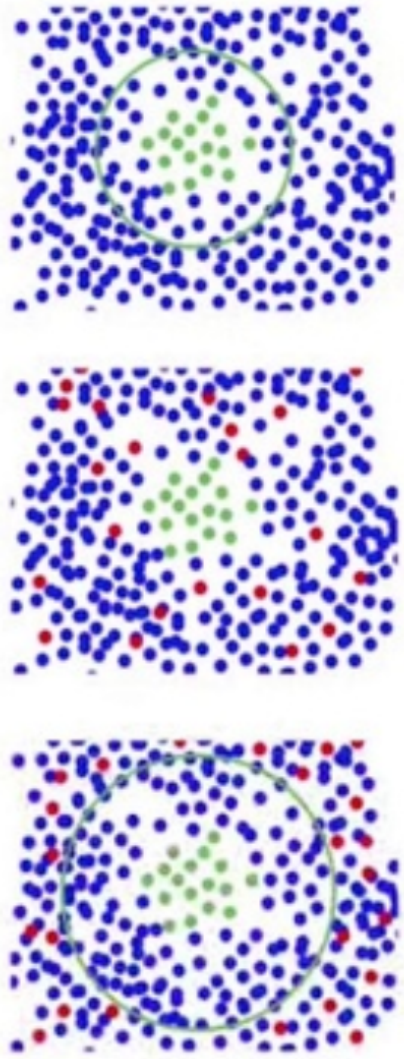
Training NMT with In-Domain vs General Domain Examples

	Medical	Law	Koran	IT	Subtitles
Medical	**56.5**	18.3	1.9	11.4	4.3
Law	21.7	**59**	2.7	13.1	5.4
Koran	0.1	0.2	15.9	0.2	0.5
IT	14.9	9.6	2.8	**43**	8.6
Subtitles	7.9	5.5	6.4	8.5	27.3
All	53.3	57.2	**20.9**	42.1	**27.6**

- Training and testing on the in-domain examples usually gives the best results
- More data is not necessary better!

Data Selection Methods

- Two methods for data selection given a small seed in-domain examples
- **Domain-Cosine**: compute the centroid of the in-domain data. Then, select the nearest-k examples
- **Domain-Finetune**: Fine-tune a pretrained LM for binary classification using random negative sampling. Then, select top-k output, or all predicted positive examples
- **Domain-Cosine + Domain-Finetune**: first run Domain-Cosine, sample negative negative examples for Domain-Finetune



Data Selection: Results

	Medical	Law	Koran	IT	Subtitles	Average
Random-500k	49.8	53.3	18.5	37.5	25.5	36.92
Moore-Lewis-Top-500k	55	58	21.4	42.7	27.3	40.88
Domain-Cosine-Top-500k	52.7	58	**22**	42.5	27.1	40.46
Domain-Finetune-Top-500k	54.8	58.8	21.8	**43.5**	27.4	**41.26**
Domain-Finetune-Positive	55.3	58.7	19.2	42.5	27	40.54
Oracle	**56.5**	**59**	15.9	43	27.3	40.34
All	53.3	57.2	20.9	42.1	**27.6**	40.22

Thank you !