Cold-start Active Learning through Self-supervised Language Modeling

Michelle Yuan[1] Hsuan-Tien Lin[2] Jordan Boyd-Graber[1]

[1]University of Maryland

[2]National Taiwan University

EMNLP 2020

Active Learning

- **Goal:** Recognize most relevant examples and query their labels from an all-knowing oracle

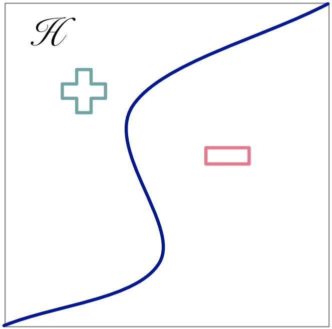
Active Learning

- **Goal:** Recognize most relevant examples and query their labels from an all-knowing oracle
- **Issue:** Traditional active learning works poorly for modern neural networks, especially during *cold-start*

Active Learning

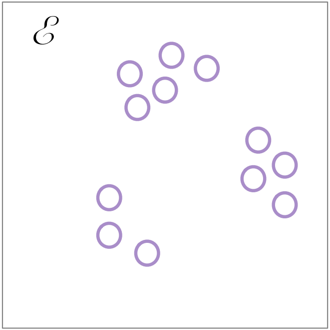
- **Goal:** Recognize most relevant examples and query their labels from an all-knowing oracle
- **Issue:** Traditional active learning works poorly for modern neural networks, especially during *cold-start*
- Limitations in SOTA NLP show a greater need for active learning *and* make active learning more difficult to deploy

Uncertainty–Diversity Dichotomy



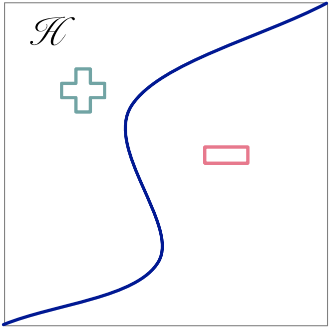
Dasgupta (2011)

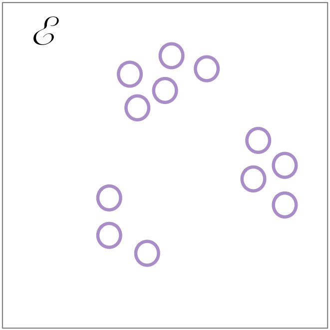
Uncertainty–Diversity Dichotomy



Dasgupta (2011)

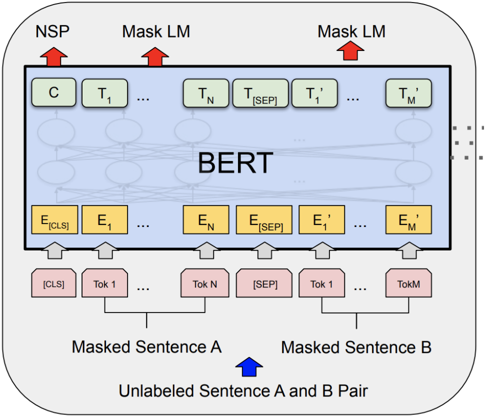
Uncertainty–Diversity Dichotomy





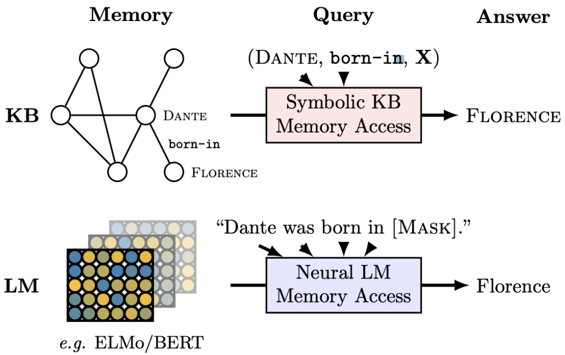
Dasgupta (2011)

Language Model Pre-training



Devlin et al. (2019)

Language Model Pre-training



Petroni et al. (2019)

ALPS

Active **L**earning by **P**rocessing **S**urprisal

ALPS

Active **L**earning by **P**rocessing **S**urprisal

1. For each sentence x in unlabeled dataset $\mathcal{U}$, compute a surprisal embedding s_x using pre-trained BERT

ALPS

Active **L**earning by **P**rocessing **S**urprisal

1. For each sentence x in unlabeled dataset $\mathcal{U}$, compute a surprisal embedding s_x using pre-trained BERT
2. Run k-means clustering on the surprisal embeddings

ALPS

Active **L**earning by **P**rocessing **S**urprisal

1. For each sentence x in unlabeled dataset $\mathcal{U}$, compute a surprisal embedding s_x using pre-trained BERT
2. Run k-means clustering on the surprisal embeddings
3. Find the sentences that are closest to each cluster center

ALPS

Active **L**earning by **P**rocessing **S**urprisal

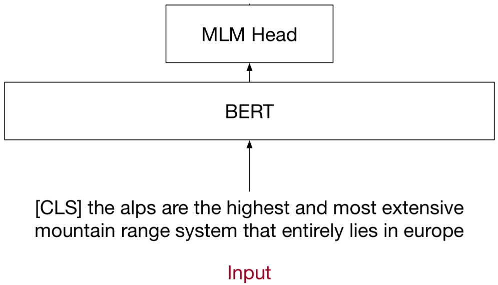
1. For each sentence x in unlabeled dataset $\mathcal{U}$, compute a surprisal embedding s_x using pre-trained BERT
2. Run k-means clustering on the surprisal embeddings
3. Find the sentences that are closest to each cluster center
4. Query labels for these k sentences

Surprisal Embeddings

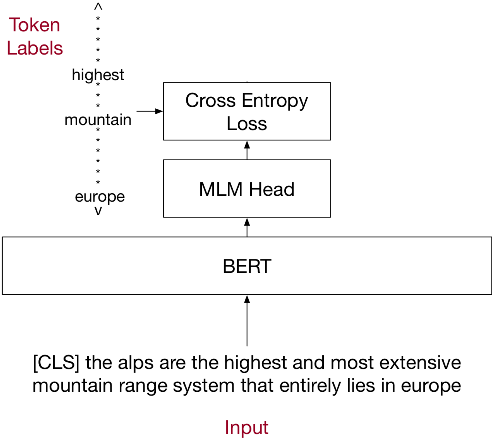
[CLS] the alps are the highest and most extensive mountain range system that entirely lies in europe

Input

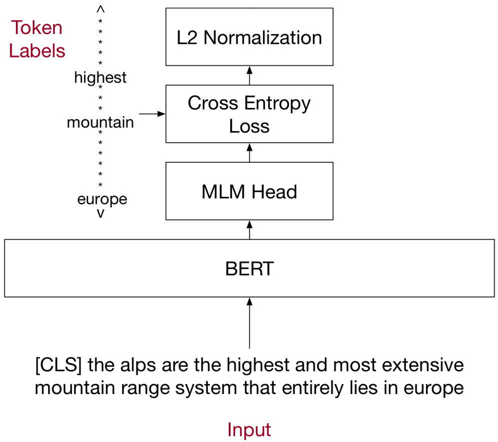
Surprisal Embeddings



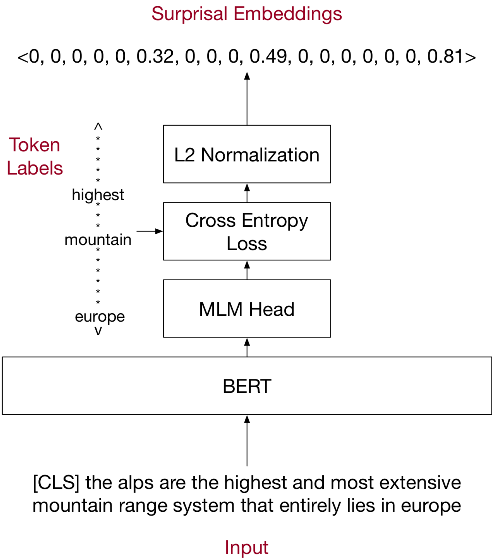
Surprisal Embeddings



Surprisal Embeddings



Surprisal Embeddings



ALPS in Action

```
Emotional eating is associated with
overeating...(background)

Ticagrelor and clopidogrel antiplatelet treatment
were used...(methods)

Visual acuity improvements in the 2 groups were
similar.  (results)

Teacher-rated and self-rated antisocial
behavior...(results)

In contrast, early intervention with selective
high-risk samples...(conclusions)

...
```

ALPS in Action

Emotional eating is associated with overeating...(background)

Ticagrelor and clopidogrel antiplatelet treatment were used...(methods)

Visual acuity improvements in the 2 groups were similar. (results)

Teacher-rated and self-rated antisocial behavior...(results)

In contrast, early intervention with selective high-risk samples...(conclusions)

...

ALPS in Action

$\langle$	0.000	0.000	0.000	0.000	0.021	0.000	0.000	... $\rangle$
$\langle$	0.043	0.000	0.000	0.000	0.000	0.385	0.000	... $\rangle$
$\langle$	0.000	0.000	0.000	0.002	0.000	0.000	0.000	... $\rangle$
$\langle$	0.000	0.000	0.000	0.000	0.000	0.000	0.039	... $\rangle$
$\langle$	0.000	0.001	0.000	0.000	0.000	0.000	0.022	... $\rangle$

...

ALPS in Action

$$\langle \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.021 \quad 0.000 \quad 0.000 \quad \ldots \quad \rangle$$

$$\langle \quad 0.043 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.385 \quad 0.000 \quad \ldots \quad \rangle$$

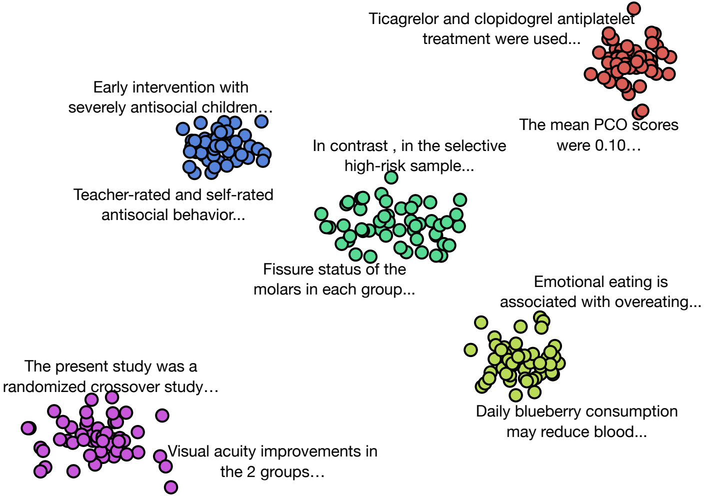
$$\langle \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.002 \quad 0.000 \quad 0.000 \quad 0.000 \quad \ldots \quad \rangle$$

$$\langle \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.039 \quad \ldots \quad \rangle$$

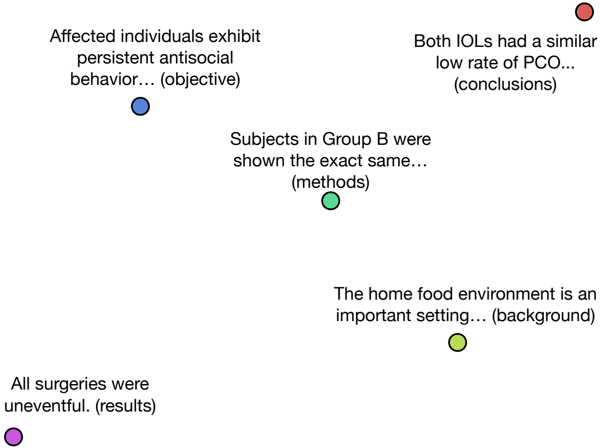
$$\langle \quad 0.000 \quad 0.001 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.000 \quad 0.022 \quad \ldots \quad \rangle$$

$$\ldots$$

ALPS in Action



ALPS in Action



Experiments

- **Task:** PUBMED 20k RCT (Dernoncourt and Lee, 2017)
- **Model:** SCIBERT (Beltagy et al., 2019)
- Simulate active learner for 10 iterations where 100 sentences are sampled each time

Baselines

1. Random

Baselines

1. Random
2. Entropy (Lewis and Gale, 1994)

Baselines

1. Random
2. Entropy (Lewis and Gale, 1994)
3. BADGE (Ash et al., 2020)

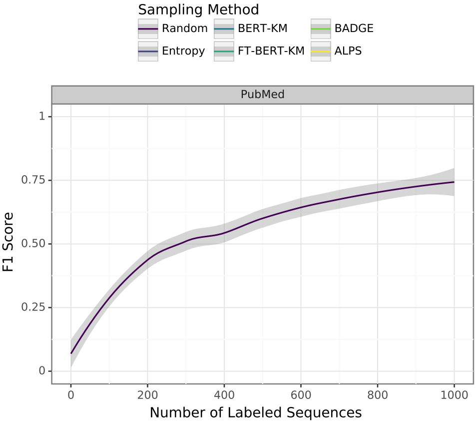
Baselines

1. Random
2. Entropy (Lewis and Gale, 1994)
3. BADGE (Ash et al., 2020)
4. BERT-KM

Baselines

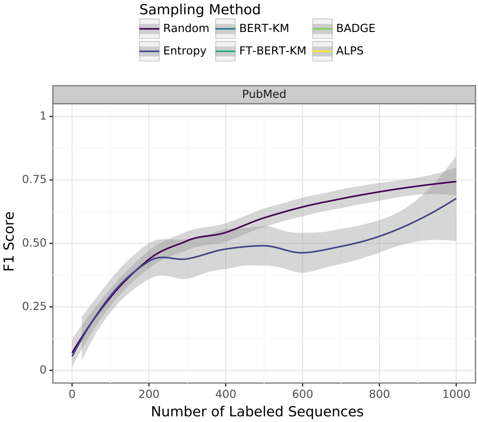
1. Random
2. Entropy (Lewis and Gale, 1994)
3. BADGE (Ash et al., 2020)
4. BERT-KM
5. FT-BERT-KM

Results



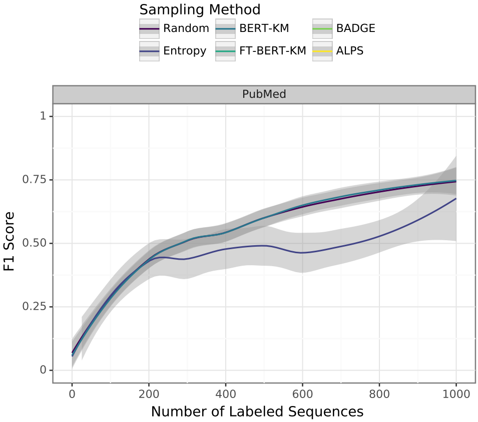
Random

Results



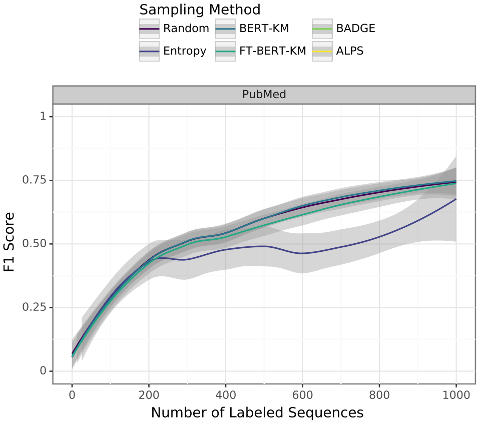
Entropy

Results



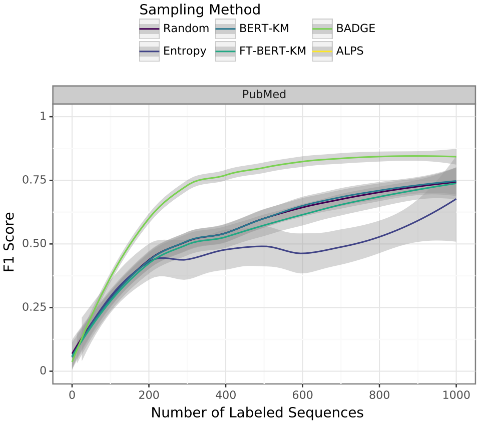
BERT-KM

Results



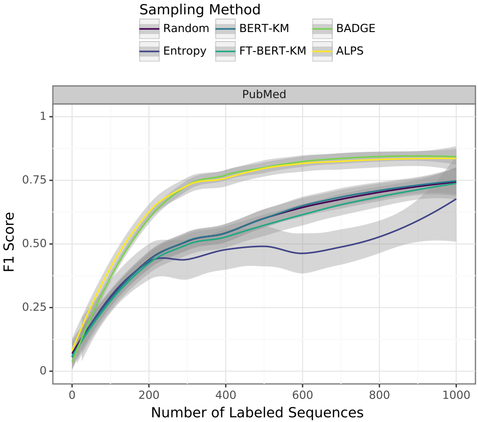
FT-BERT-KM

Results



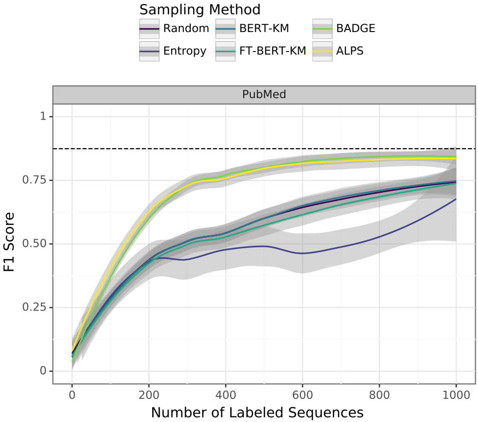
BADGE

Results



ALPS

Results



Full training dataset

Time

	AG NEWS	PUBMED
Random	< 1	< 1
Entropy	7	10
ALPS	14	24
BADGE	23	70
BERT-KM	28	58
FT-BERT-KM	33	79

References I

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP*.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. *IJCNLP*, 2:308–313.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP*.