

Industry Scale Semi-Supervised Learning for Natural Language Understanding

집현전 최신반

3조 : 박동주, 이정훈, 이인환,
조소영

발표자 : 박동주

Industry Scale Semi-Supervised Learning for Natural Language Understanding

Luoxin Chen*

Alexa AI

luoxchen@amazon.com

Francisco Garcia*

Alexa AI

fgmz@amazon.com

Varun Kumar*

Alexa AI

kuvrun@amazon.com

He Xie*

Alexa AI

hexie@amazon.com

Jianhua Lu

Alexa AI

jianhual@amazon.com

NAACL2021 Industry track

Paper: <https://www.aclweb.org/anthology/2021.naacl-industry.39.pdf>

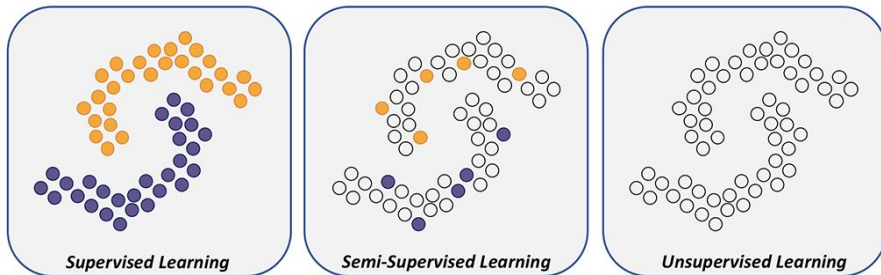
Voice-Assistants

- Voice-assistants with speech and natural language understanding (NLU) are becoming increasingly prevalent in everyday life.
- An NLU system commonly consists of an intent classifier (IC) and named entity recognizer (NER).
- For example, if a user asks “play lady gaga”, the IC classifies the query to intent of PlayMusic, and the NER classifies “lady gaga” as Artist.



Semi-Supervised Learning

- Semi-Supervised Learning (SSL) provides a framework for utilizing large amount of unlabeled data when obtaining labels is expensive
- A common practice to evaluate SSL algorithms is to take an existing labeled dataset and only use a small fraction of training data as labeled data, while treating the rest of the data as unlabeled dataset.



Challenges while applying SSL techniques at scale including

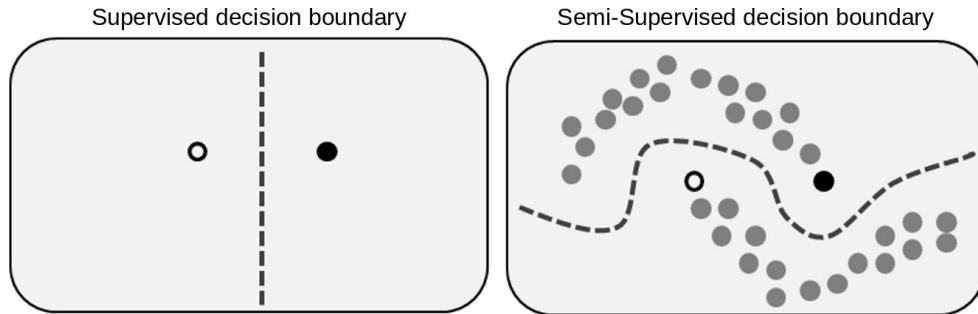
- How much unlabeled data should we use for SSL and how to select unlabeled data from a large pool of unlabeled data?
- Most SSL benchmarks make the assumption that unlabeled datasets come from the same distribution as the labeled datasets.
- Unlike widely used NLU datasets such as SNIPS, ATIS, real-world voice assistant datasets are much larger and have a lot of redundancy because some queries such as “turn on lights” might be much more frequent than others.

Contribution

- Design of a production SSL pipeline which can be used to intelligently select unlabeled data to train SSL models
- Experimental comparison of four SSL techniques including, Pseudo-Label, Knowledge Distillation, Cross-View Training, and Virtual Adversarial Training in a real-world setting using data from Amazon Alexa.
- Operational recommendations for NLP practitioners who would like to employ SSL in production setting.

Semi-Supervised Learning

- In supervised learning, given a labeled dataset DI composed of input-label pairs (x, y) , the goal is to learn a prediction model $f_{\theta}(x)$, with parameters θ , that is able to predict the correct label y' corresponding to a new unseen input instance x' .
- SSL techniques aim to leverage an unlabeled dataset, Du , to create better performing models than those that could be obtained by only using DI .



Semi-Supervised Learning

- In this paper, we conduct comprehensive experiments and analysis related to these commonly used SSL techniques, and discuss their pros and cons in the industry setting.
- Model confidence-based data selection is a widely used technique for SSL data selection where unlabeled data is selected on the basis of a classifier's confidence.
- Due to the abundance of unlabeled data in production voice assistants, model confidence-based filtering leads to a very large data pool. To overcome this issue, we study different data selection algorithms which can further reduce the size of unlabeled data.

Research Question

- We are interested in studying two different questions relevant to the use of unlabeled data in production environments.
 - How to effectively select SSL data from a large pool of unlabeled data.
 - how do SSL techniques perform in realistic scenarios?
- To do so, we focus on the tasks of intent classification (IC) and named entity recognition (NER), two important components in NLU systems.

Model Architecture

- LSTM-based multi-task model for IC and NER tasks
- 300-dimension fastText word embeddings, trained on a large voice assistant corpus.
- Shared 256-dimension Bi-LSTM encoder and two separate task-specific Bi-LSTM encoders.
- Softmax layer and a conditional random field layer.

Data Selection Approaches

- In the industry setting, we often encounter the situation where we have extremely large pool of unlabeled data, intractable to have SSL methods run on the entire dataset.
- Given this challenge, we propose a two stage data selection pipeline to create an unlabeled SSL pool, D_u , of a practical size, from the much larger pool of available data.

Data Selection Approaches

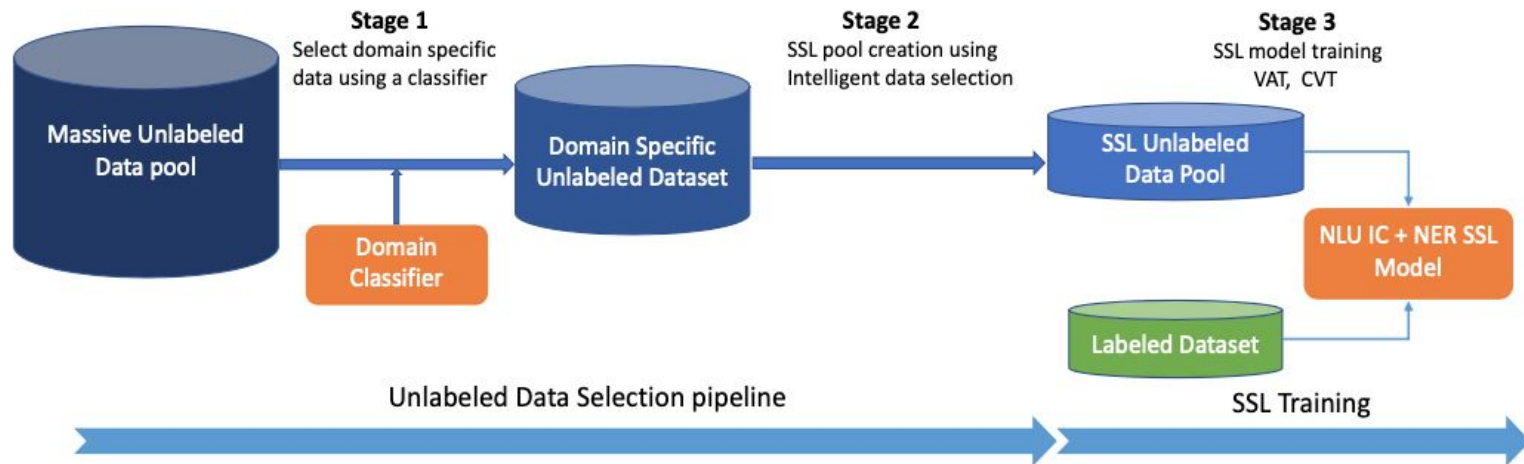


Figure 1: SSL pipeline. Domain specific unlabeled data are first selected using a domain classifier. We then select a subset of the unlabeled data using submodular optimization or committee based selection. Finally we train different SSL models using selected data combined with the labeled data.

Data Selection Approaches – Selection by Submodular Optimization

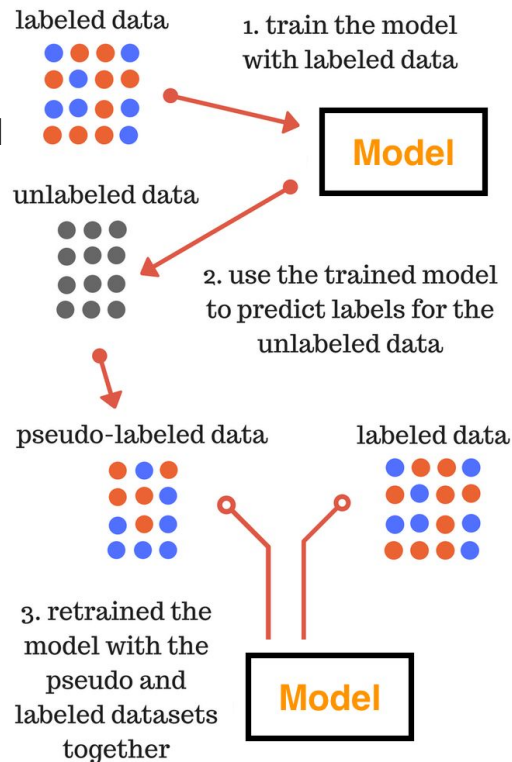
- The method proposed in **Submodularity for Data Selection in Machine Translation**
- Use feature-based submodular selection, where submodular functions are given by weighted sums of non-decreasing concave functions applied to modular functions.
- For SSL data selection, we use 1–4 n-gram as features and logarithm as the concave function.
- Filter out any n-gram features which appear less than 30 times in $D_I \cup D_u$.
- The algorithm starts with D_I as the selected data and chooses the utterance from the candidate pool D_u which provides maximum marginal gain.

Data Selection Approaches – Selection by Committee

- Data points with high uncertainty are more likely to be incorrectly predicted than those with low uncertainty.
- To detect data points on which the model is not reliable, we train a committee of n teacher models (we use $n = 4$ in this paper), and compute the average entropy of the probability distribution for every data point.
- Identify an entropy threshold with an acceptable error rate for mis-annotations (e.g., 20%) based on a held-out dataset.

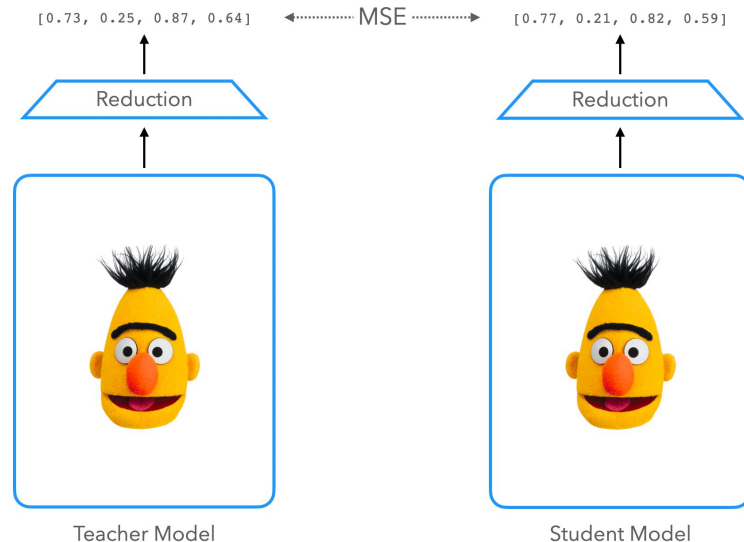
Semi-Supervised Learning – Pseudo Label (PL)

- In Pseudo Label, a teacher model trained on labeled data is used to produce pseudo-labels for the unlabeled data set.
- A student model trained on the union of the labeled and pseudo-labeled data sets, often outperforms the teacher model.

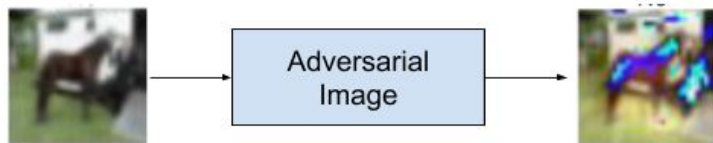


Semi-Supervised Learning – Knowledge Distillation (KD)

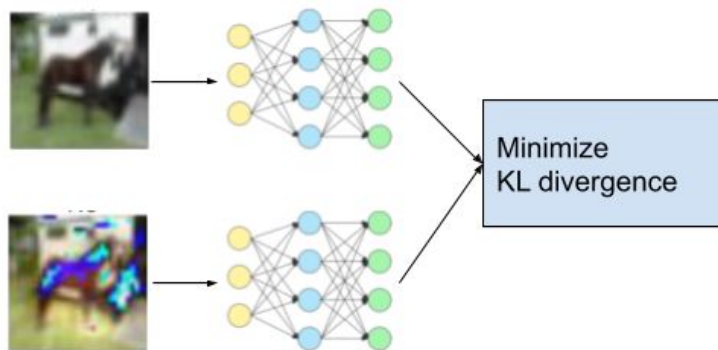
- On the other hand, KD SSL methods do not assign a particular label to an unlabeled instance, but instead consider the whole distribution over the label space
- In KD, it is hypothesized that leveraging the probability distribution over all labels provides more information than assuming a definitive label belonging to one particular class.






Semi-Supervised Learning – Virtual Adversarial Training

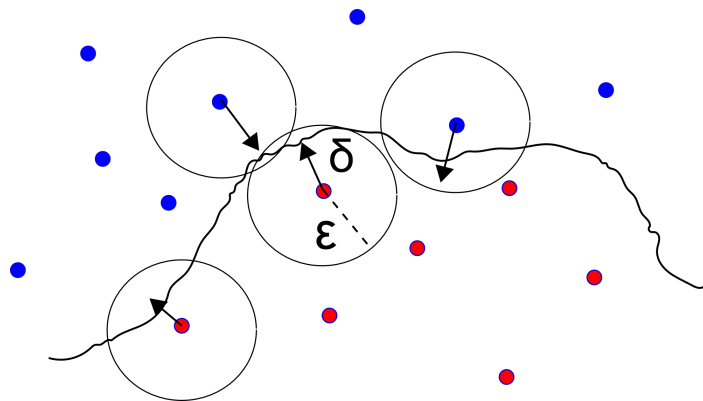


Step 1: Generate the adversarial image



Step 2: Minimize the KL divergence

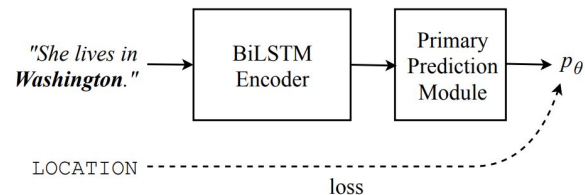
	$+ .007 \times$		$=$	
x		$\text{sign}(\nabla_x J(\theta, x, y))$		$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
"panda"		"nematode"		"gibbon"
57.7% confidence		8.2% confidence		99.3 % confidence



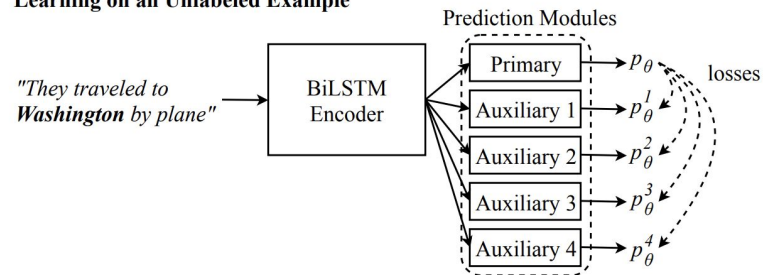
Semi-Supervised Learning – Cross-View Training

- The model is trained with standard supervised learning on labeled examples.
- On unlabeled examples, auxiliary prediction modules with different views of the input are trained to agree with the primary prediction module.

Learning on a Labeled Example



Learning on an Unlabeled Example



Inputs Seen by Auxiliary Prediction Modules

- Auxiliary 1: They traveled to _____
Auxiliary 2: They traveled to *Washington* _____
Auxiliary 3: _____ *Washington* by plane
Auxiliary 4: _____ by plane

Commercial Dataset

- Labeled training data and unlabeled data come from a similar distribution.
- For each domain, our dataset contains 50k unique training, 50k unique testing utterances.
- 500K unlabeled data pool (Stage 1) → 300k unlabeled data pool (Stage 2)

Table 1: Relative error rate reduction using KD, over baseline trained with only labeled data, for Music domain. Unlabeled data SSL pool size varies from 50K to 1M utterances. 50K labeled examples are used for all experiments. The metric for IC is classification error rate, and for NER is entity recognition F1 error rate.

Task	50K	100K	300K	500K	1M
IC	-3.81%	-3.37%	-4.40%	-4.49%	-4.09%
NER	-6.05%	-7.49%	-6.96%	-8.07%	-7.20%

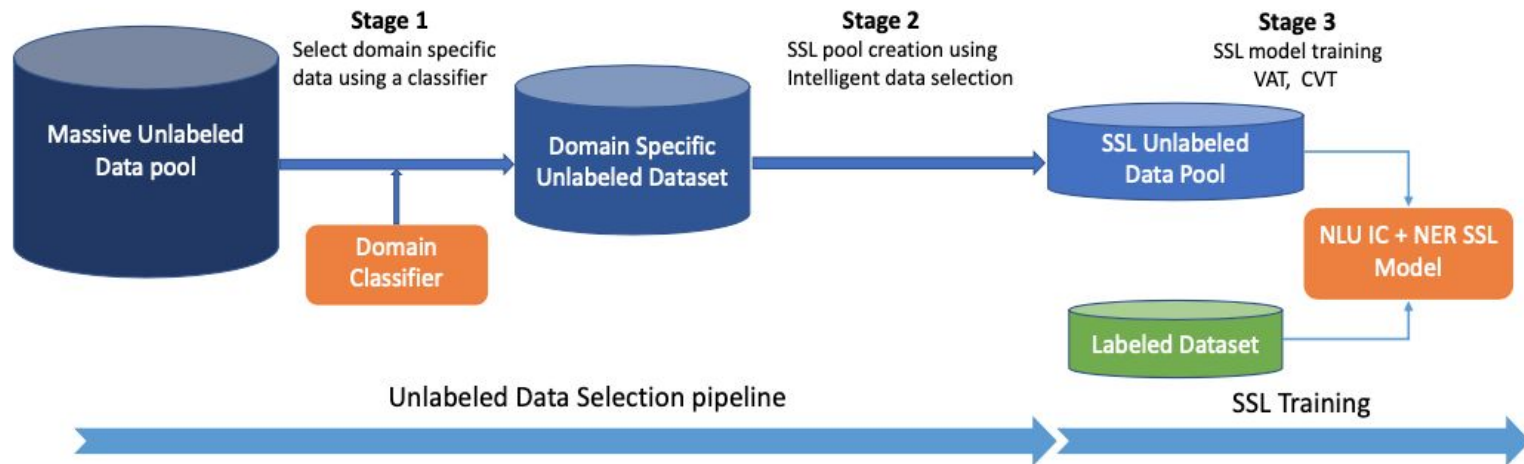


Figure 1: SSL pipeline. Domain specific unlabeled data are first selected using a domain classifier. We then select a subset of the unlabeled data using submodular optimization or committee based selection. Finally we train different SSL models using selected data combined with the labeled data.

SNIPS Dataset

- Labeled and unlabeled data come from different distributions.
- Labeled data pool from SNIPS dataset
- Unlabeled data pool from commercial dataset
- 300K unlabeled data pool (Stage 1) → 20k unlabeled data pool (Stage 2)

Results on Commercial Dataset

Table 2: Error reduction of SSL methods, relative to baseline. **Bold** represents the best SSL method for a given data selection technique. **Bold[†]** represents the best performance across all SSL methods and data selection techniques.

SSL Algorithm	Selection Approach	Communication		Music		Notifications		ToDos	
		IC	NER	IC	NER	IC	NER	IC	NER
Baseline	Random	0	0	0	0	0	0	0	0
PL		-3.61%	-2.86%	-4.86%	-3.70%	-2.79%	-4.06%	-2.94%	-3.33%
KD		-6.35%	-2.97%	-6.96%	-4.40%	-3.48%	-4.84%	-4.18%	-1.59%
VAT		-8.14%	-8.18%	-11.15%	-9.26%	-6.90%	-8.55%	-4.07%	-4.59%
CVT		-9.61%	-5.26%	-7.21%	-8.13%	-7.39%	-7.19%	-4.75%	-2.38%
Baseline	Submodular	0	0	0	0	0	0	0	0
PL		-4.90%	-3.11%	-4.61%	-3.35%	-1.48%	-4.62%	-1.70%	-4.32%
KD		-6.69%	-3.40%	-8.19%	-3.63%	-2.91%	-4.32%	-5.01%	-2.59%
VAT		-11.56%	-8.39%	-14.72% [†]	-11.03% [†]	-8.70%	-11.86% [†]	-6.24%	-5.77%
CVT		-14.72%	-5.91%	-9.84%	-9.94%	-8.72% [†]	-10.61%	-6.30%	-3.13%
Baseline	Committee	0	0	0	0	0	0	0	0
PL		-10.54%	-3.91%	-9.02%	-3.93%	-6.90%	-4.47%	-4.55%	-3.67%
KD		-11.13%	-4.46%	-11.98%	-4.09%	-7.76%	-5.06%	-6.10%	-2.61%
VAT		-13.16%	-9.40% [†]	-13.63%	-10.10%	-8.50%	-11.82%	-5.75%	-5.99% [†]
CVT		-15.25% [†]	-6.53%	-8.72%	-8.27%	-8.72% [†]	-10.40%	-7.34% [†]	-3.58%

Results on SNIPS Dataset

Table 3: Model performance by different SSL methods and data selection methods, for SNIPS data set. The metric for IC task is classification error rate, and for NER task is entity recognition F1 error rate.

SSL Algorithm	Selection Approach	SNIPS	
		IC	NER
Baseline	Random	0.9744	0.9367
PL		0.9743	0.9326
KD		0.9743	0.9424
VAT		0.9814	0.9604
CVT		0.9871	0.9565
Baseline	Submodular	0.9744	0.9367
PL		0.9743	0.9342
KD		0.9786	0.9403
VAT		0.9728	0.9579
CVT		0.9785	0.9524
Baseline	Committee	0.9744	0.9367
PL		0.9700	0.9272
KD		0.9729	0.9353
VAT		0.9772	0.9501
CVT		0.9780	0.9518

Diversity of Selected Data

- Measure the diversity of the selected data by computing the unique n-gram ratio present in $\mathcal{D}_l \cup \mathcal{D}_u$ and \mathcal{D}_l data.
- We observe that a diverse SSL pool does not necessarily lead to better performance.
- This result highlights that simply optimizing for token diversity is not enough for improving SSL performance.

Table 4: Unique unigram and 1-4 grams ratio present in $\mathcal{D}_l \cup \mathcal{D}_u$ and \mathcal{D}_l

Domains	Random		Committee		Submod	
	Unigram	1-4 gram	Unigram	1-4 gram	Unigram	1-4 gram
Communication	3.21	9.29	3.29	10.21	1.41	6.17
Todos	2.88	6.04	1.4	3.51	1.51	3.19
Music	3.19	6.42	3.24	6.39	3.43	7.18
Notifications	3.04	6.01	3.08	5.9	1.77	3.97

Recommendations & Limitation



Recommendations

- Prefer VAT and CVT SSL techniques over PL and KL.
- **Use data selection to select a subset of unlabeled data.**
 - Recommend Submodular Optimization based data selection in light of its lower cost and similar performance to committee based method.

Conclusion

- In this paper, we conduct extensive experiments and in-depth analyses of different SSL techniques applied to industry-scale NLU tasks.
- We investigate different data selection approaches including submodular optimization and committee-based filtering.
- Our paper provides insights on how to build an efficient and accurate NLU system, utilizing SSL, from different perspectives (e.g. model accuracy, amount of data, training time and cost, etc).

Thank you