Natural Language Processing
自然語言處理

黃瀚萱
Department of Computer Science
National Chengchi University
2020 Fall

Semi-Supervised Approaches to NLP

Schedule

Date	Topic
9/16	Introduction
9/23	Linguistic Essentials
9/30	Collocation
10/7	Language Model
10/14	Performance Evaluation and Word Sense Disambiguation
10/21	Text Classification (HW1 will be assigned)
10/28	Invited Talk: NLP and Cybersecurity (Term Project)
11/4	POS Tagging
11/11	Midterm Exam

Schedule

Date	Topic
11/18	Chinese Word Segmentation
11/25	Word Embeddings
12/2	**Neural Networks for NLP**
12/9	Semi-supervised Learning
12/16	**Discussion about your Final Project**
12/23	**Invited Talk**
12/30	**Discourse Analysis**
1/6	Final Project Presentation II
1/13	Final Exam

Important Dates

Date	Event
10/05	Release of Dataset Part I
10/28	**Tutorial in Class**
11/10	Release of Dataset Part II
11/18	**Submit Your Team Information to Moodle and Register**
12/13	Registration Due
12/16	**Discussion of Your Final Project (In Class)**
12/14 - 12/21	**Formal Run (Result Submission)**
12/25	Announcement of Formal Run Scores
12/31	Final Report Submission
2020/01/06	**Final Project Presentation**
2021/01/08	Announcement of Final Scores

Agenda

- Introduction
- Challenging issues of low resource NLP tasks
- Semi-supervised approaches to NLP
- Case studies

Low Resource NLP

- Many natural language processing tasks are tackled with machine learning approaches in these days.
- However, machine learning models usually require large amounts of annotated data to train, in particular, the neural networks with high expressive power.

Issue of Low Resource

- Statistical machine translator (SMT) still outperform neural machine translator (NMT) in some scenario where the parallel instances are limited.
    - Low resource languages.
    - Applications in new domains.
- Large amounts of annotated data do not exist for for many low-resource languages, and for high-resource languages it can be difficult to find linguistically annotated data of sufficient size and quality to allow neural methods to excel.

Goal

- This topic aims to bring together researchers from the NLP and ML communities who work on learning with neural methods when there is not enough data for those methods to succeed out-of-the-box.

- Techniques may include self-training, paired training, distant supervision, domain adaptation, semi-supervised and transfer learning as well as , and human-in-the-loop techniques such as active learning.

Low Resource NLP Tasks

- New NLP tasks suffer from lack of labeled data because of their nature of novelty and complexity.
  - Complex tasks require large amount of training data for machine learning models.

Reasonable Labeled Data Size

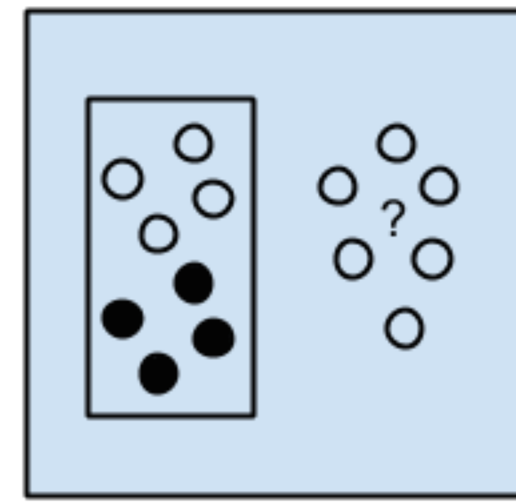
Type	Task	Reasonable Data Size
Sentence Classification	Sentiment Analysis	5K~
Document Classification	News Categorization	5K~
Relation Recognition	Discourse Relation Recongition	10K~
Labeling	POS Tagging	5K~
	Chinese Word Segmentation	5K~
Parsing	Dependency Parsing	10K~
Generation	Summarization	100K~
	Machine Translation	1M~

Bottleneck of Manual Annotation

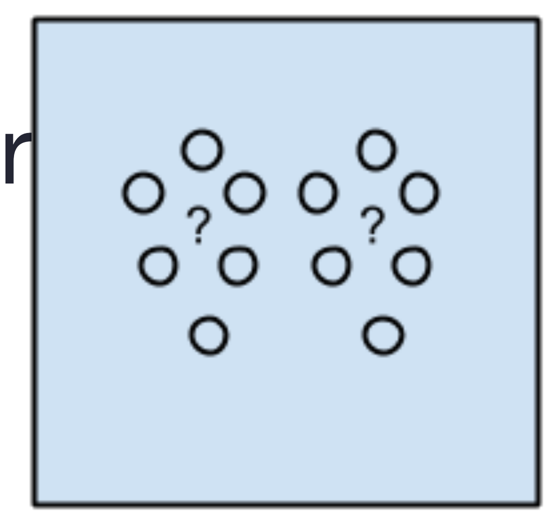
- Costly
  - Amazon Mechanical Turk does not always work.
- Time-consuming
  - Latency
- Impractical in some cases
  - The occurrences of ironic sentences in the Amazon reviews are only 0.1%!

Supervised vs. Unsupervised Learning

- Supervised learning:
  - Training on fully-labeled data
- Unsupervised learning
  - No label is available
  - Can still perform topic modeling or cluster analysis



Supervised Learning Algorithms

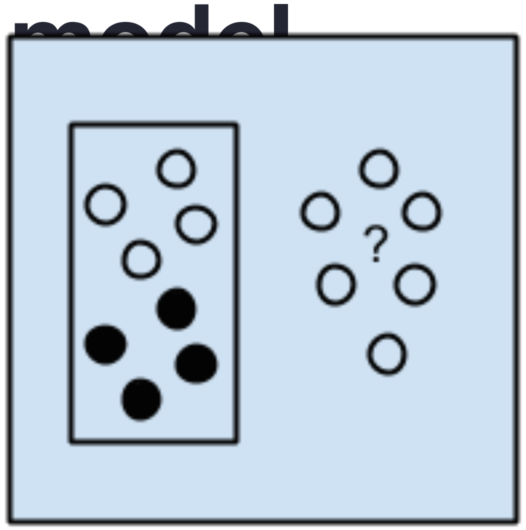


Unsupervised Learning Algorithms

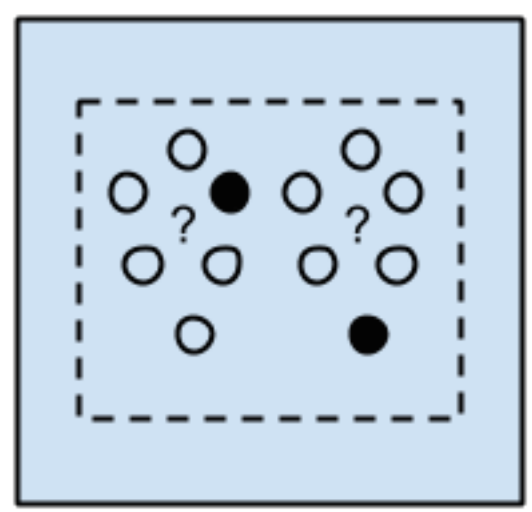
Semi-Supervised Learning

- Part of data are fully-labeled
- Data are labeled for related tasks only
- Data are pseudo-labeled

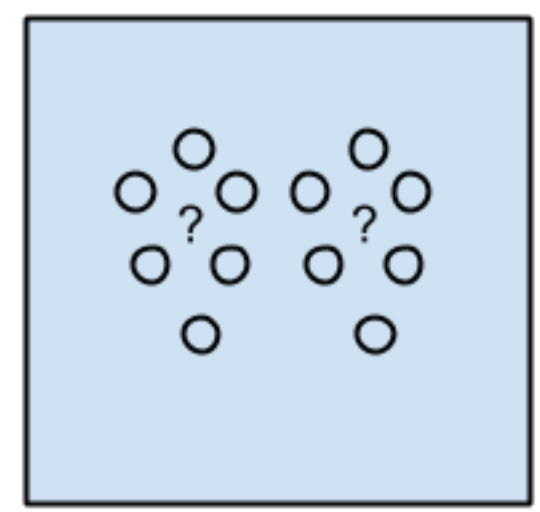
Unlabeled data provides information for improving the model!



Supervised Learning Algorithms



Semi-supervised Learning Algorithms



Unsupervised Learning Algorithms

Semi-supervised Learning

- Pseudo-labeled data
  - Self-training
  - Data augmentation
- Pre-training
- Multitask learning
- Transfer learning

Pseudo-labeled Data

- Large amount of self-labeled data is available on the Internet and usually used as training/test data for a wide range of NLP tasks.
    - 😊 => Postive sentiment
    - 😟 => Negative sentiment
- Though the self-labeled data is very useful, it may suffer from serious reliability issues.

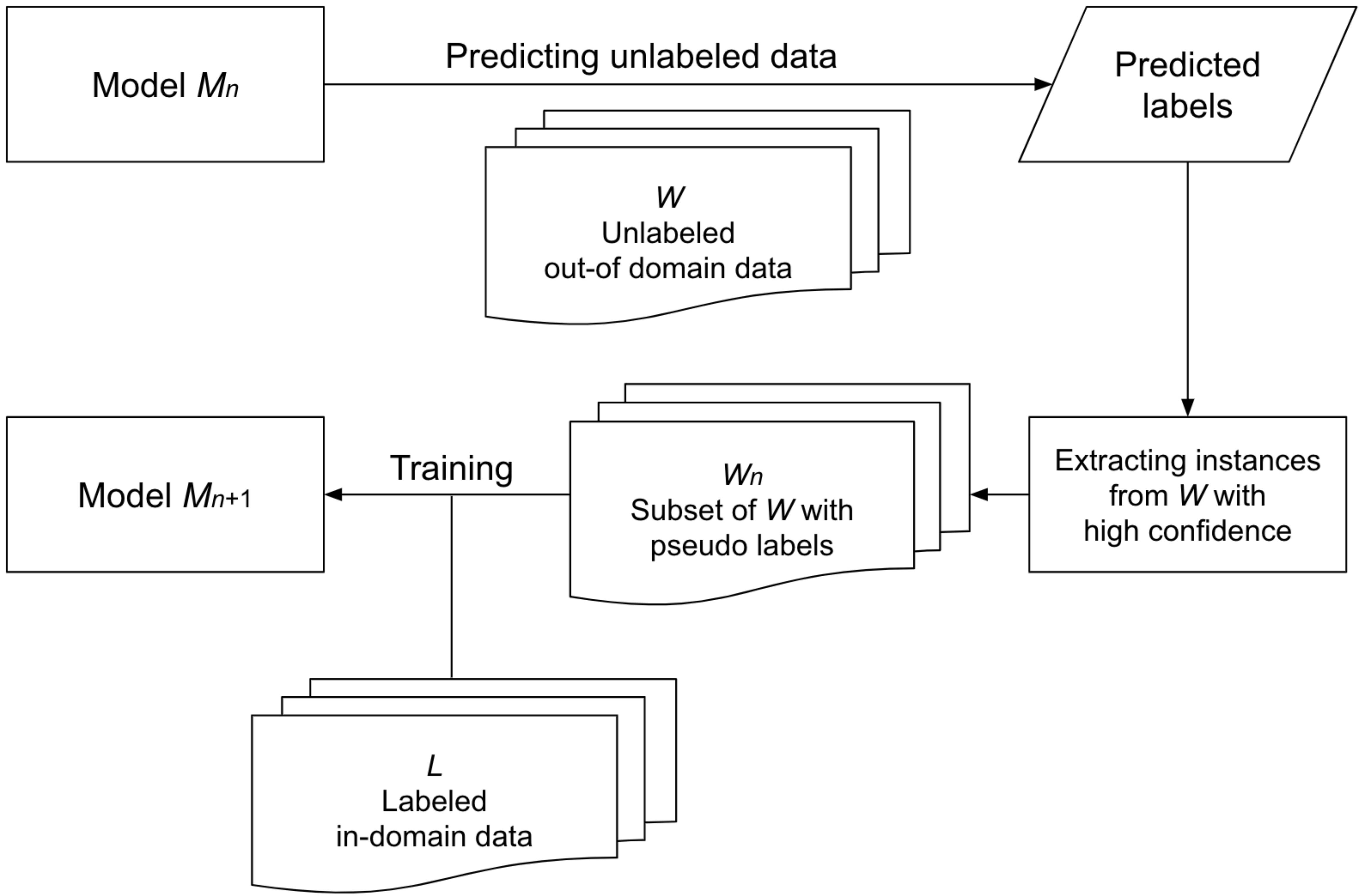
Self-training

- Train an initial model m_0 with labeled data L_0
- for i = 1 … n
  - Use $m_{i\text{-}1}$ to predict unlabeled data, label the ones with a high confidence as pseudo labeled data L_i

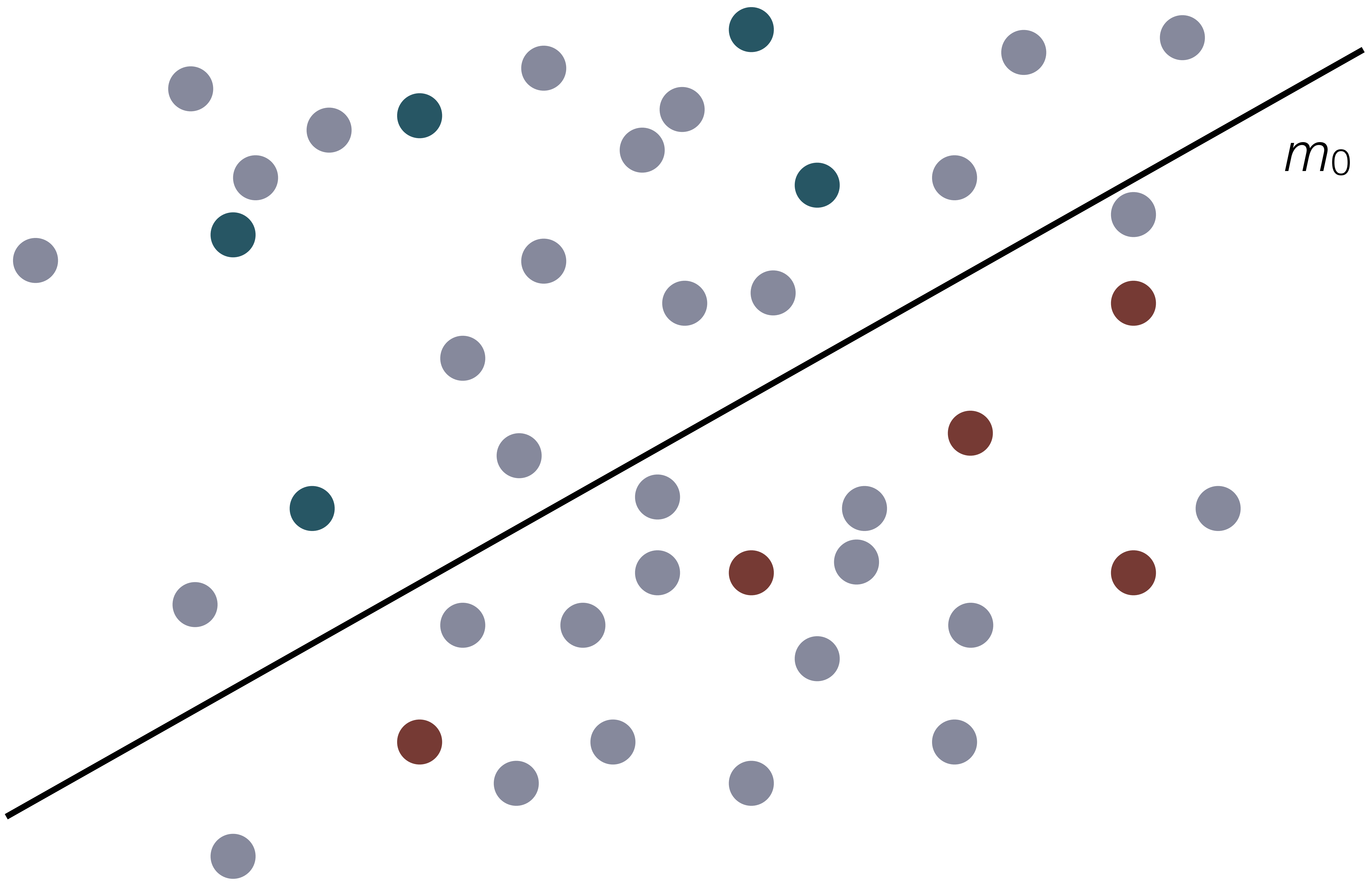
Algorithm 1 Self-training

repeat
    $m \leftarrow train\_model(L)$
    for $x \in U$ **do**
        if $\max m(x) > \tau$ **then**
            $L \leftarrow L \cup \{(x, p(x))\}$
until no more predictions are confident

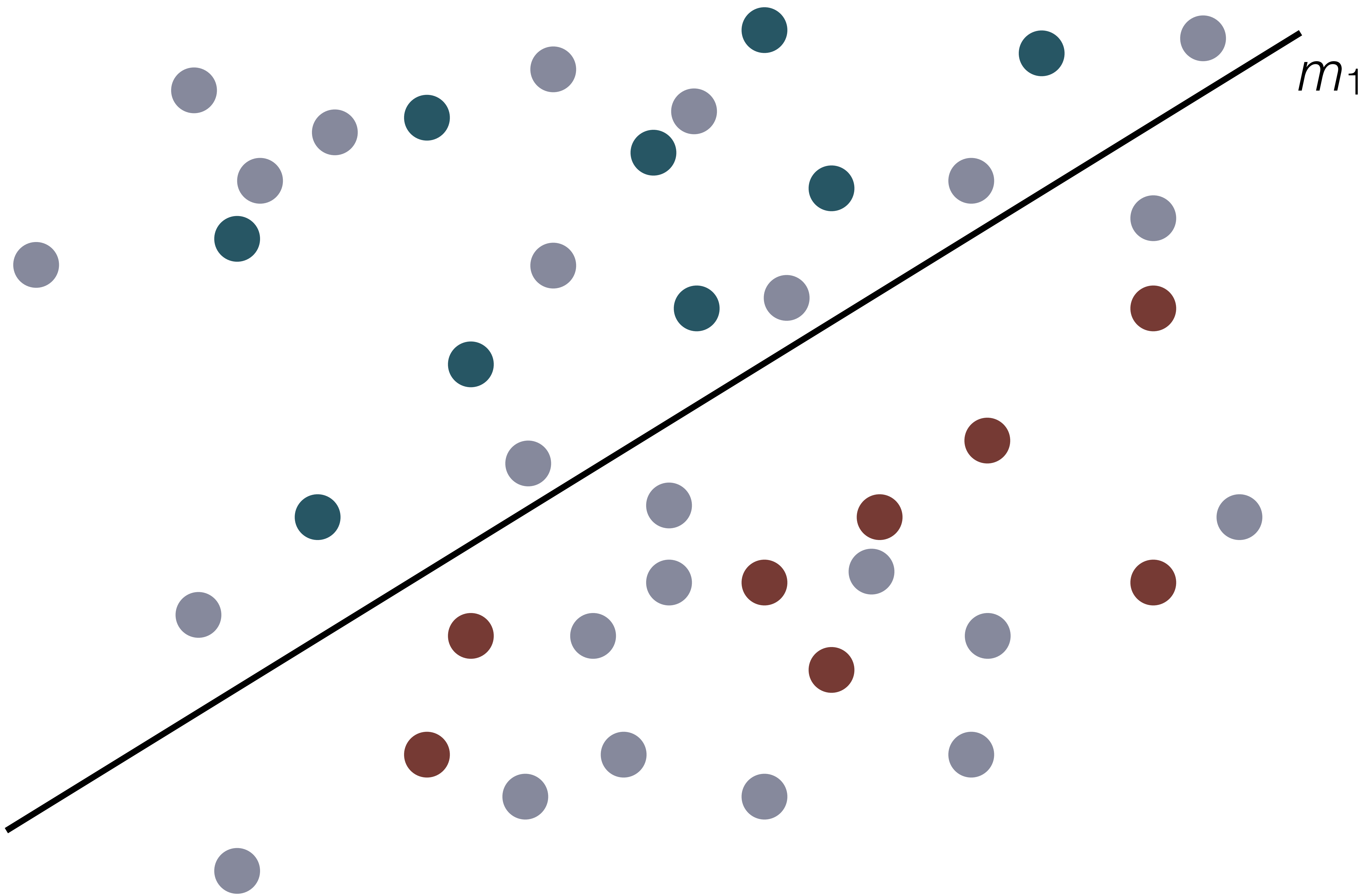
Flow-chart of Self-training



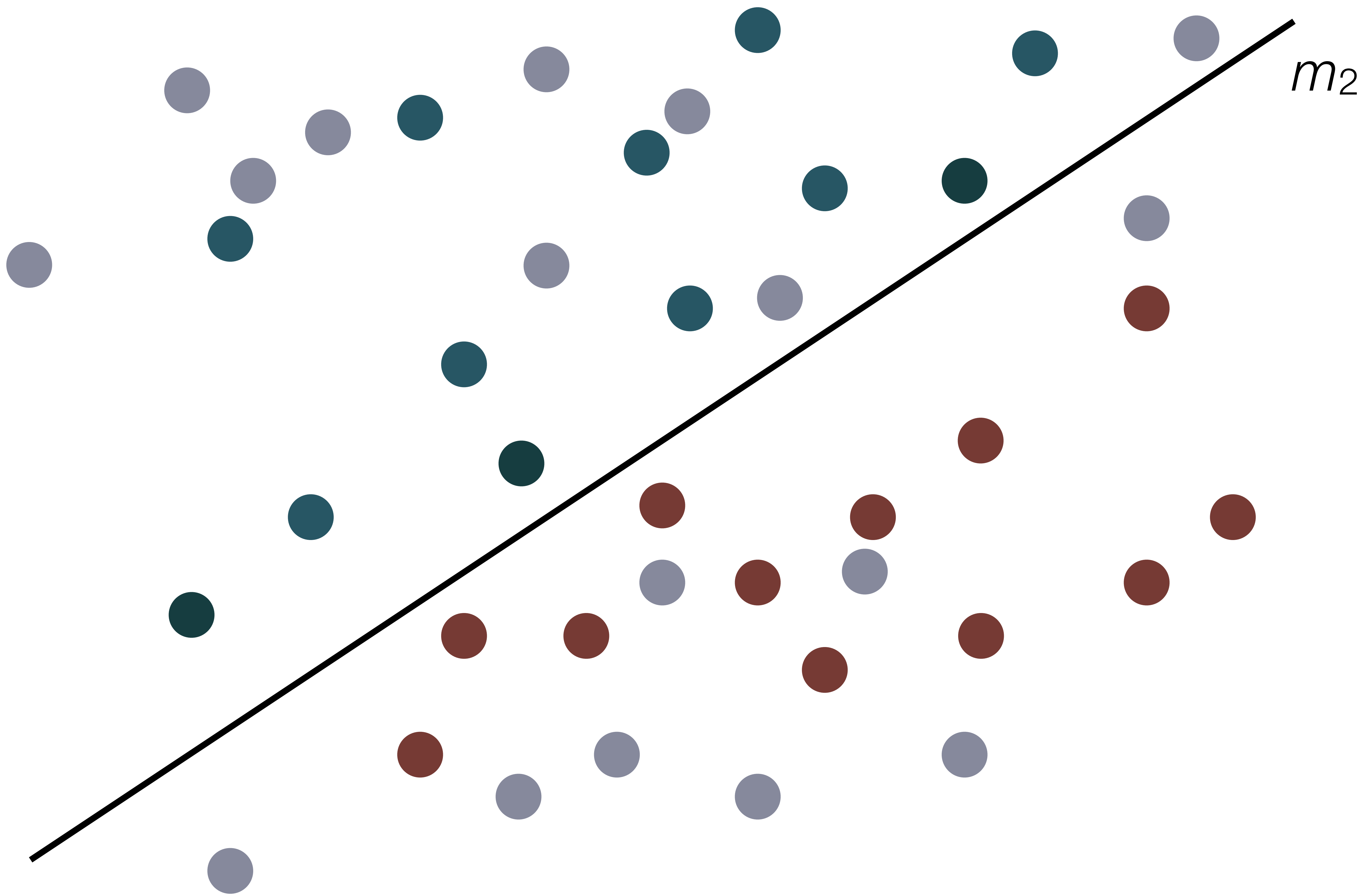
Self-training



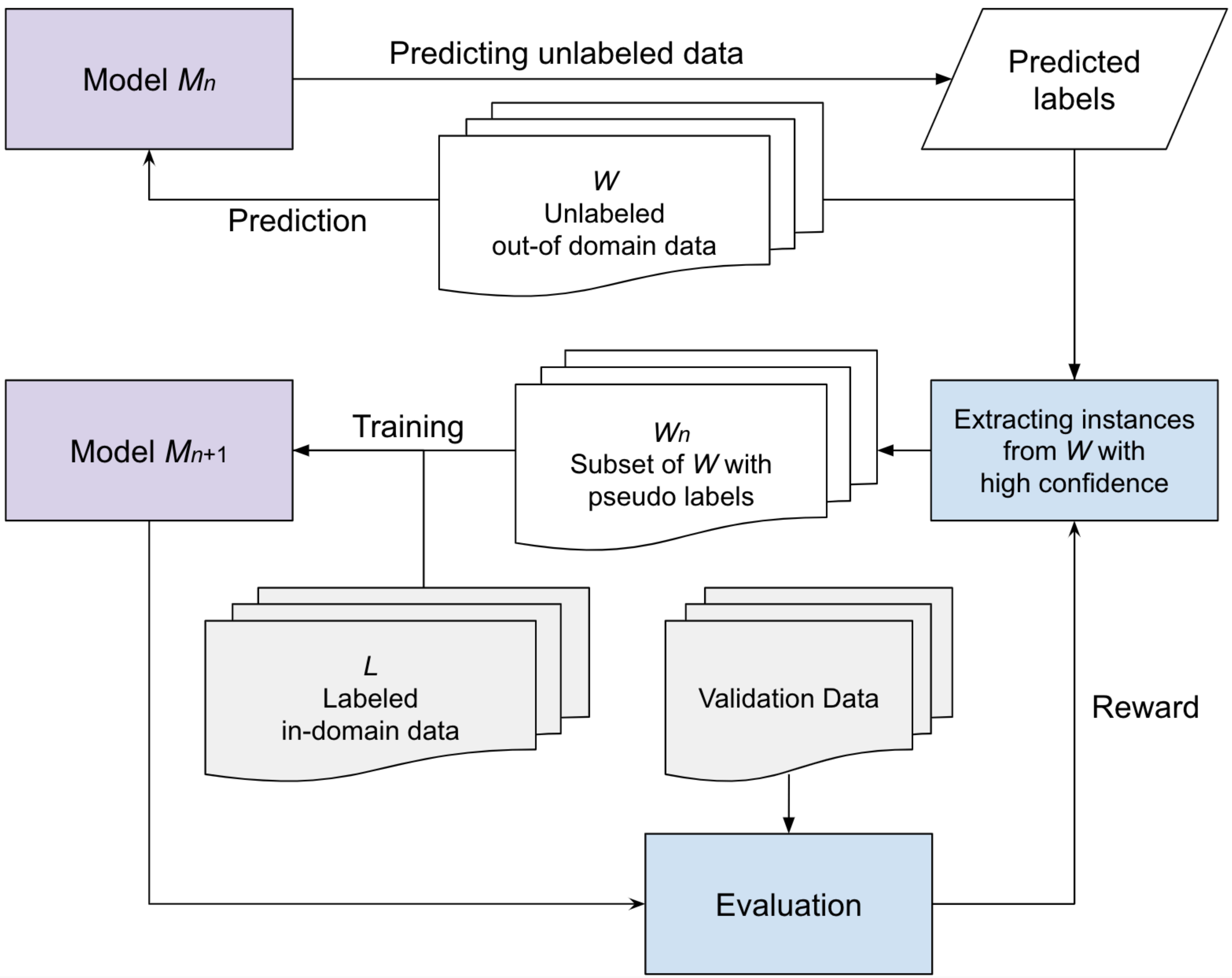
Self-training



Self-training



Self-training with Reinforcement Learning



Pre-training of Neural Networks

- Word level
  - CBOW, Skip-gram, GloVe, FastText, etc.
- Sentence level
  - ELMo, BERT, XLNet, T5, etc.
- Document level

Word Embeddings

- Word embeddings (or distributed word representations) are trained to predict well words that appear in its context.

- Given a set of sentences $w_1, ..., w_T$, the objective of the skip-gram mo [...] the log-likelihood:

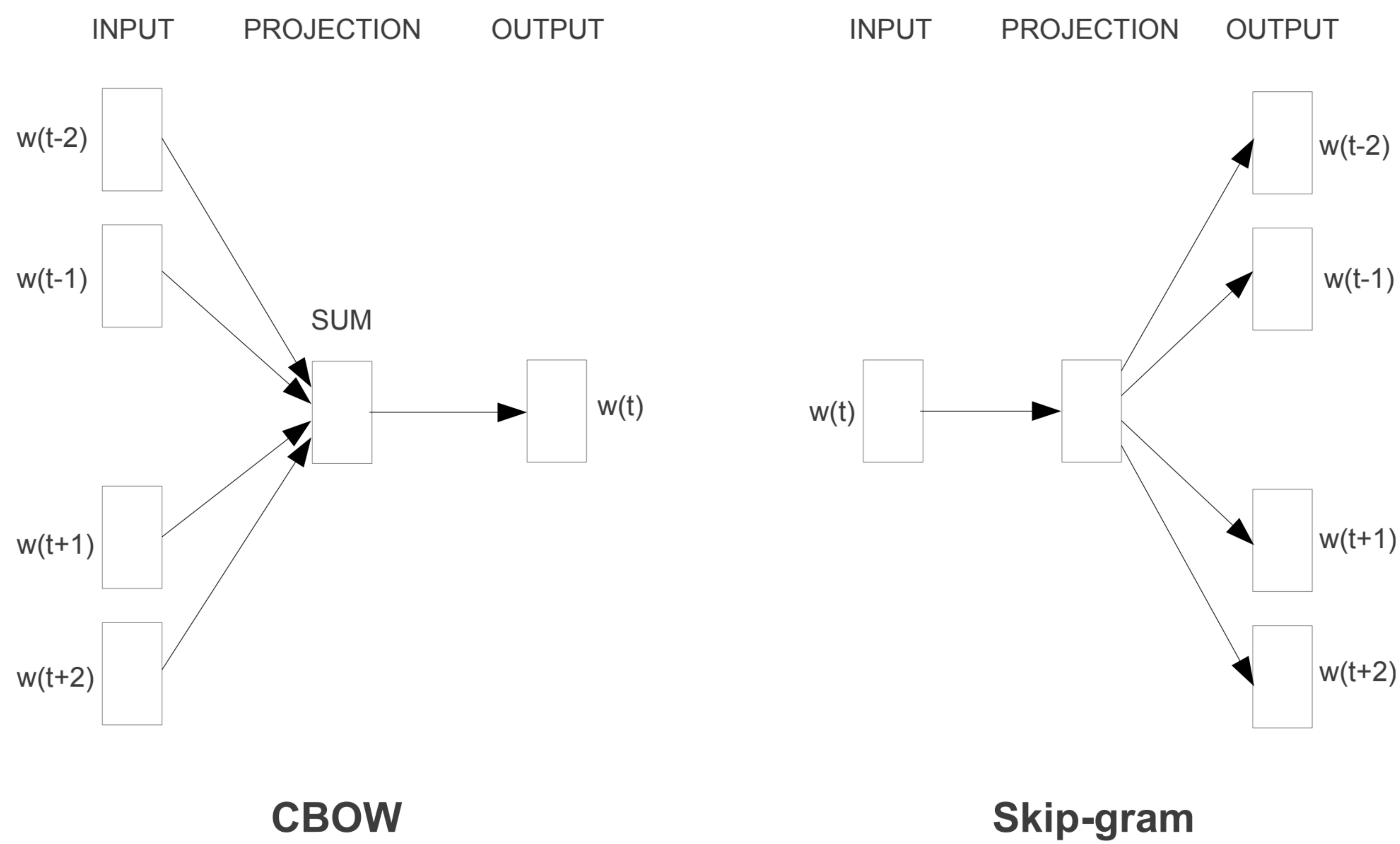
$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c \mid w_t)$$

- With a scorinc function s maps pairs of a target word and a context [...]:

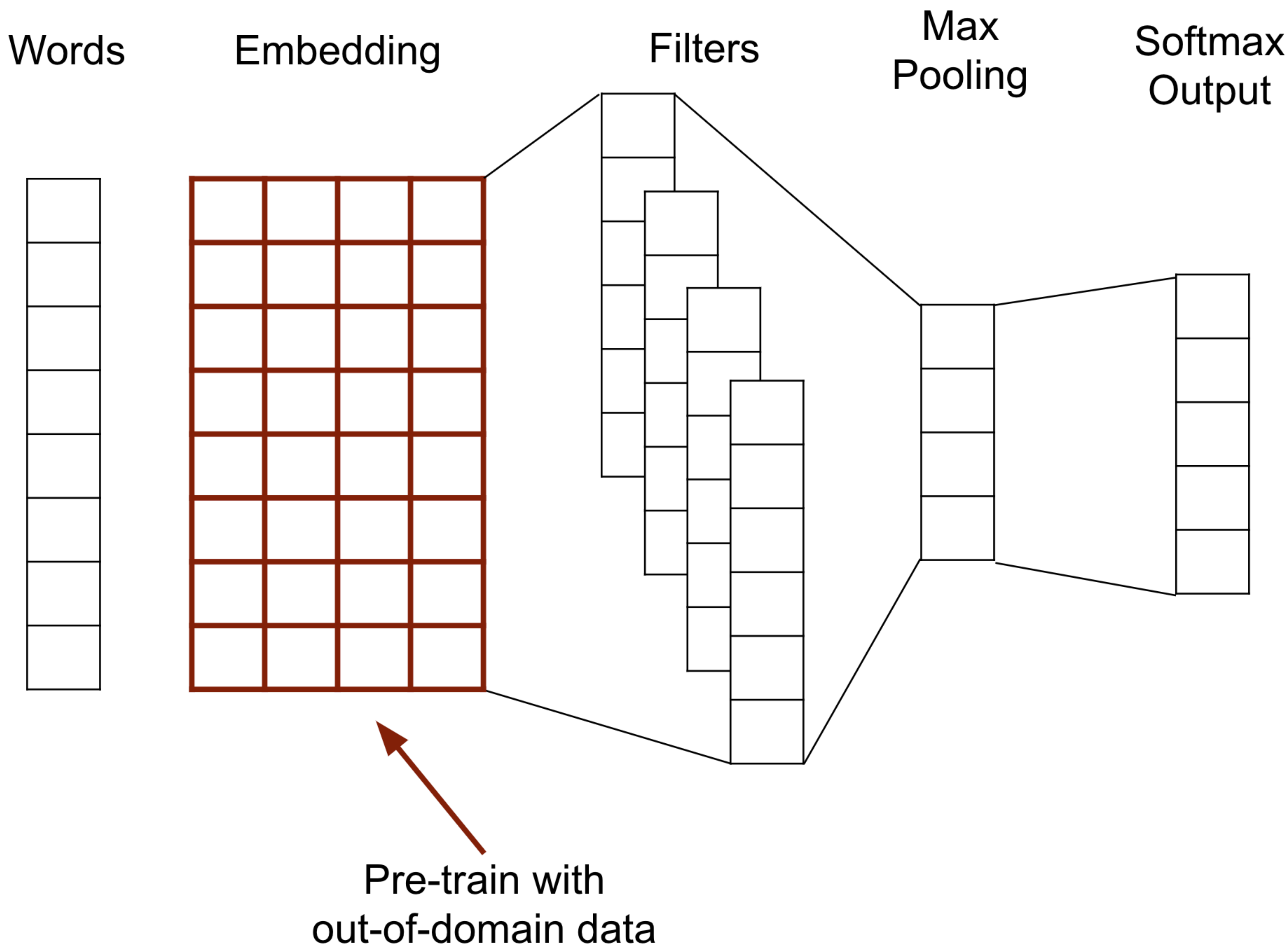
$$p(w_c \mid w_t) = \frac{e^{s(w_t,\, w_c)}}{\sum_{j=1}^{W} e^{s(w_t,\, j)}}$$

CBOW vs Skip-gram

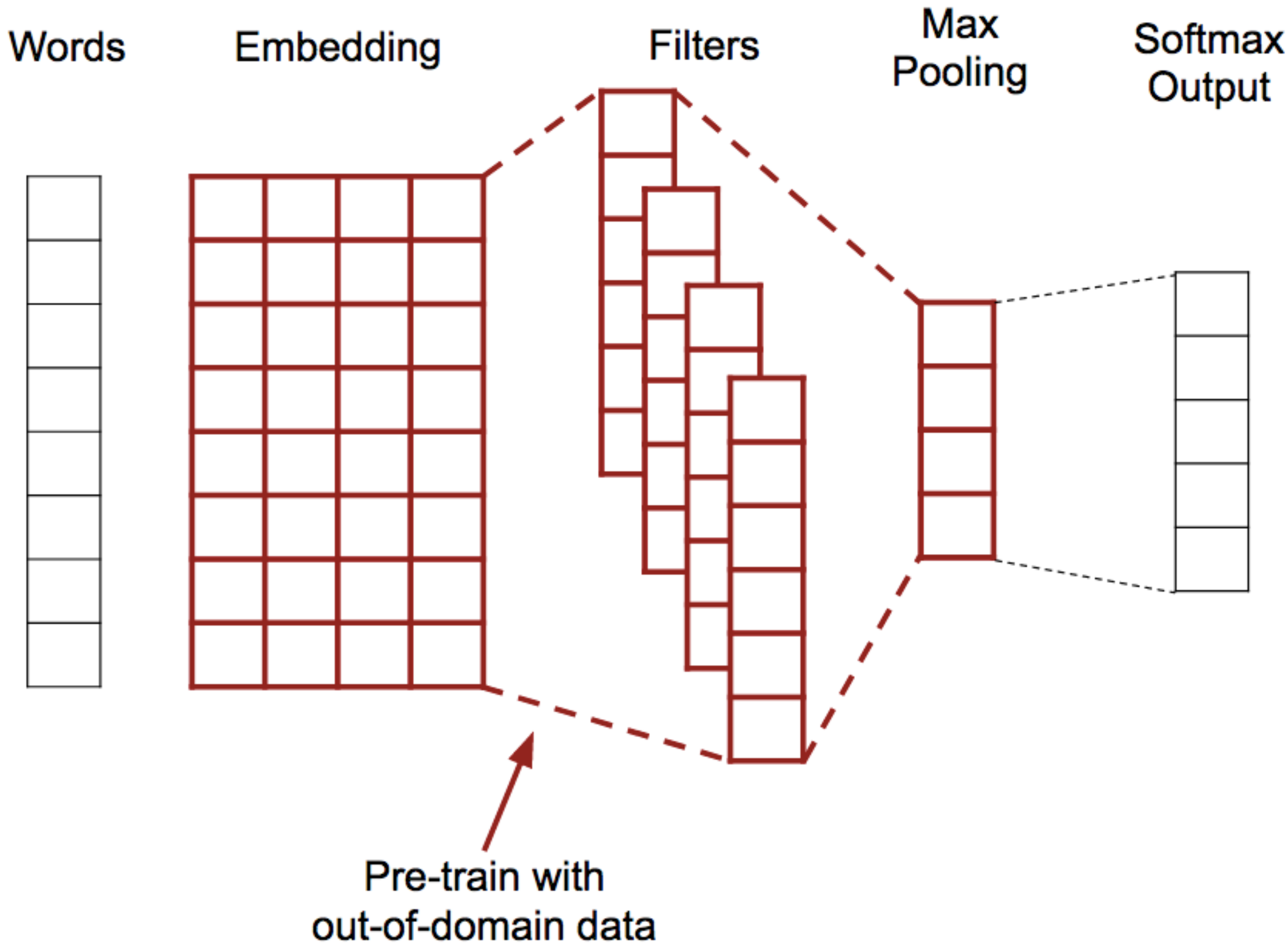
- CBO ontext.
- Skip-
  curre



Word Embeddings as Pre-trained



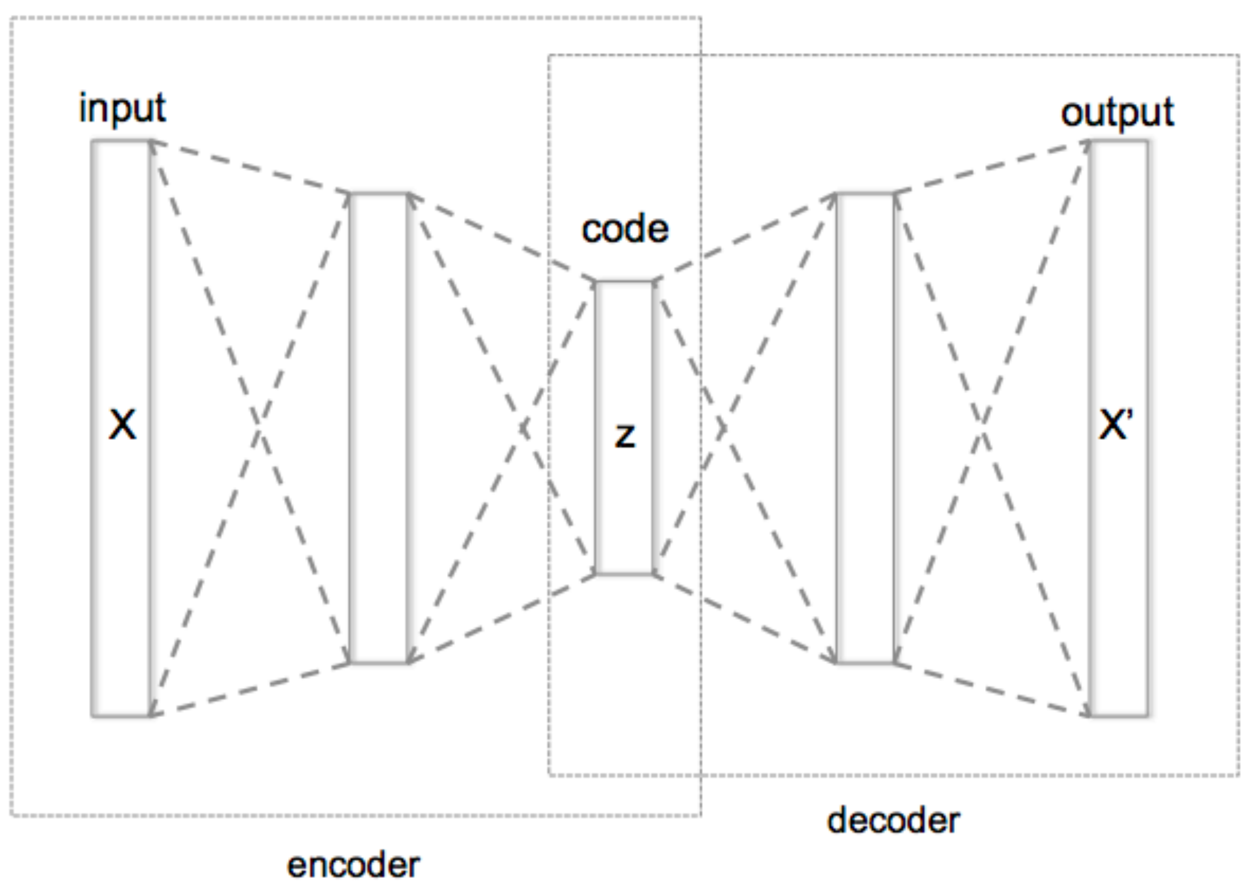
Pre-training the Sentence Representation



Auto Encoder

- A large set of sentences can be used to train an auto encoder in the unsupervised manner.

- The encoder can be used as a sentence embedding model.



As similar t
the input as

Data Augmentation

We had an amazing night at the hotel ↔ **Positive**

Paraphrase generation

We **spent** a **wonderful** night **in** the hotel.

We had a **fantastic** night at the hotel

We had a **great** night at the hotel

We had a **great** night in the hotel

Back Translation

Genuine in English	We had an amazing night at the hotel
Translated to Chinese	我們在酒店度過了一個美好的夜晚。
Back Translated to English	We spent a wonderful night in the hotel.

Downside of Back Translation

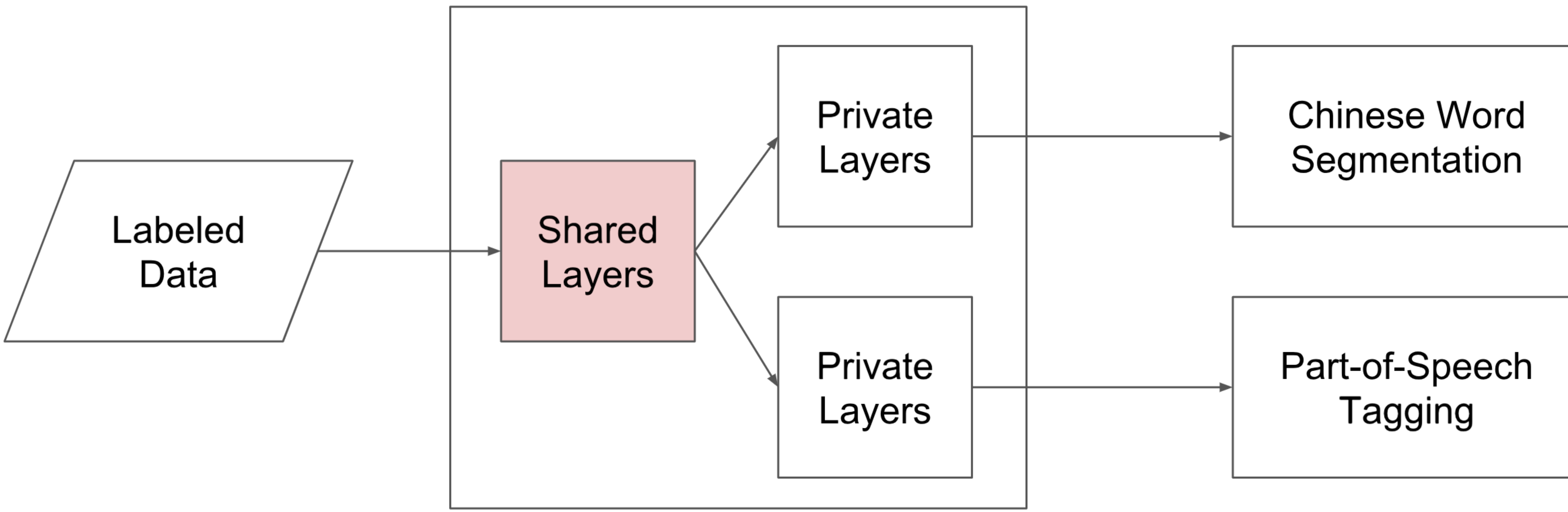
- The quality is highly relied on the machine translation model.
- The powerful online MT cannot be used for privacy data.
- Unsuitable for some tasks
  - Some aspects of linguistic phenomena will be lost after translation
    - Hate language, offensive language, dirty words, etc.

Co-Training

- Multitask learning, in which related tasks with large amount of data are introduced to co-train the main task NN model, is a popular approach for improving the main task.
    - The auxiliary task can be an unsupervised one or other tasks with a lot of training data.
- Adversarial learning can further applied for transferring source domain knowledge to target domain.

Multi-task Learning

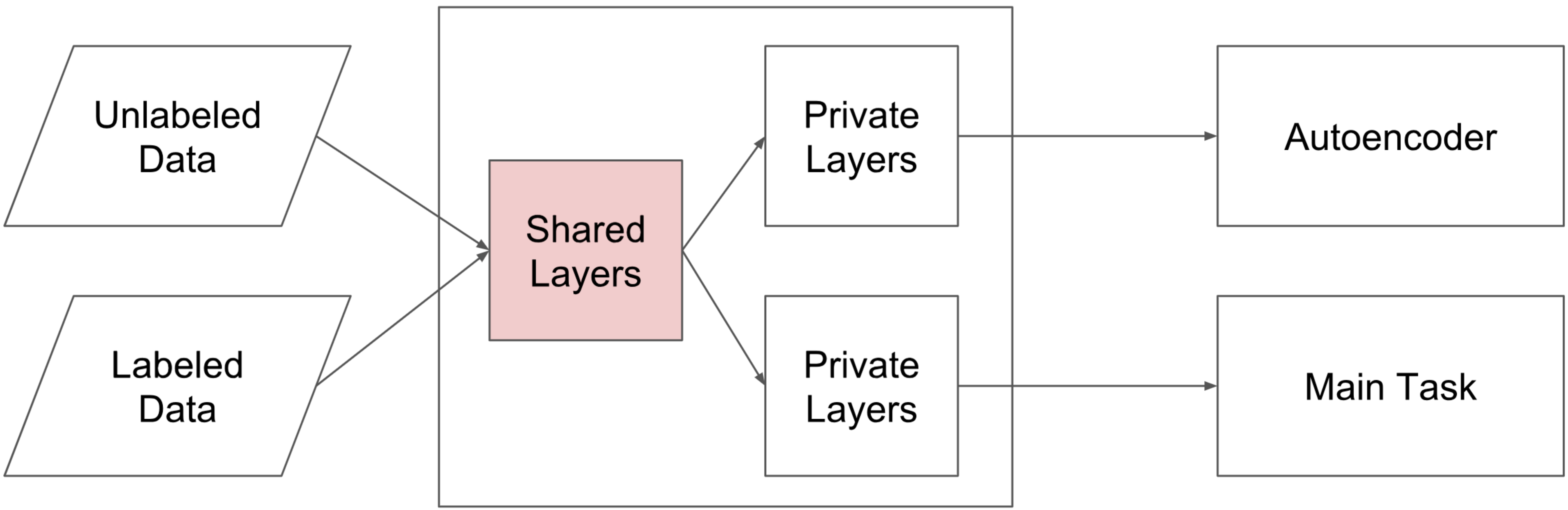
- Train a better representation with large amount of data in related tasks.



$$Loss = \lambda * l_{main} + (1-\lambda) * l_{auxiliary}$$

Multi-task Learning

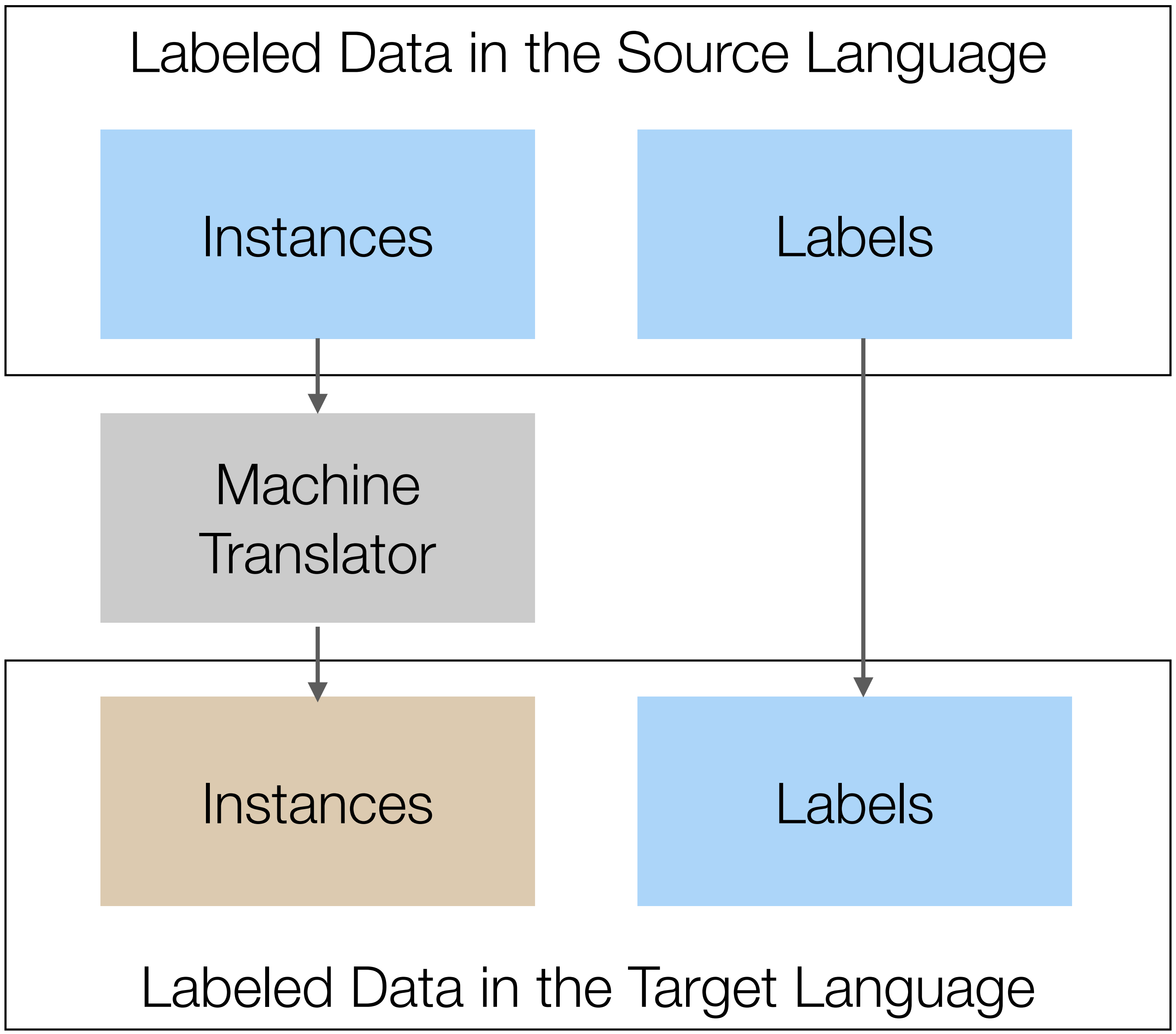
- Or even in an unsupervised setting.



Cross-lingual Transfer Learning

- We have a lot of training data in the source language
  - Many datasets are available for English
- And we would to build a model for deal with the data in the target language
  - Less datasets are available for Chinese and other languages

Cross-lingual Transfer Learning with MT



Machine Translation

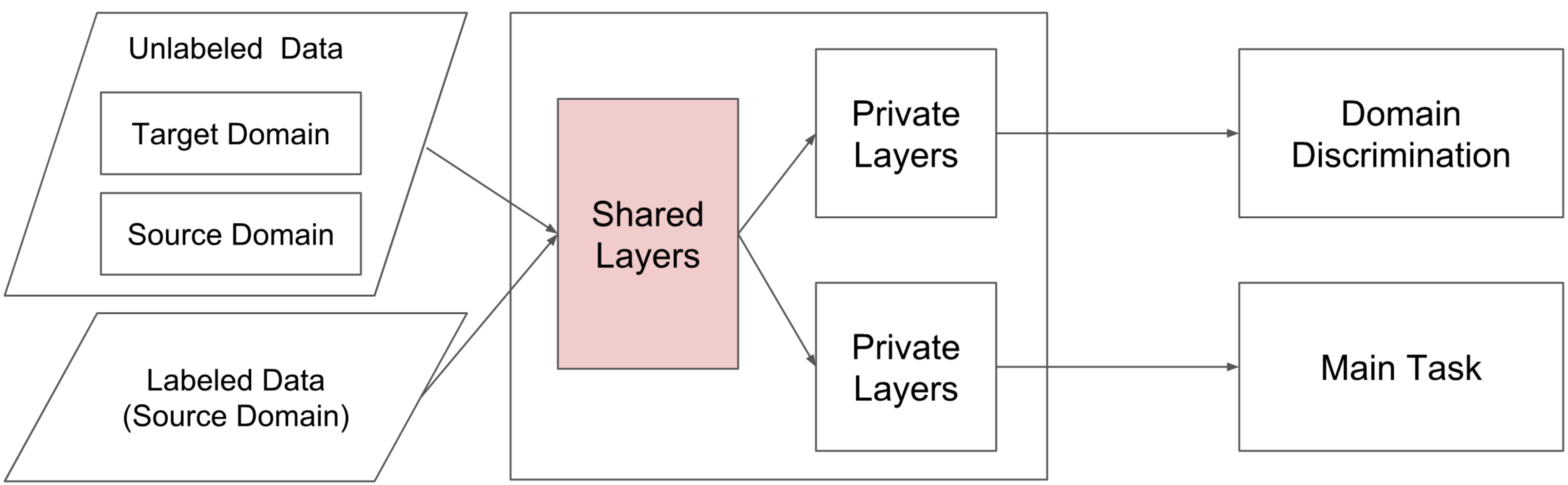
- The latest machine translator can generate very good translated training or testing data.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

Adversarial Learning

- Optimization ensures the domain discriminator cannot distinguish the domain.

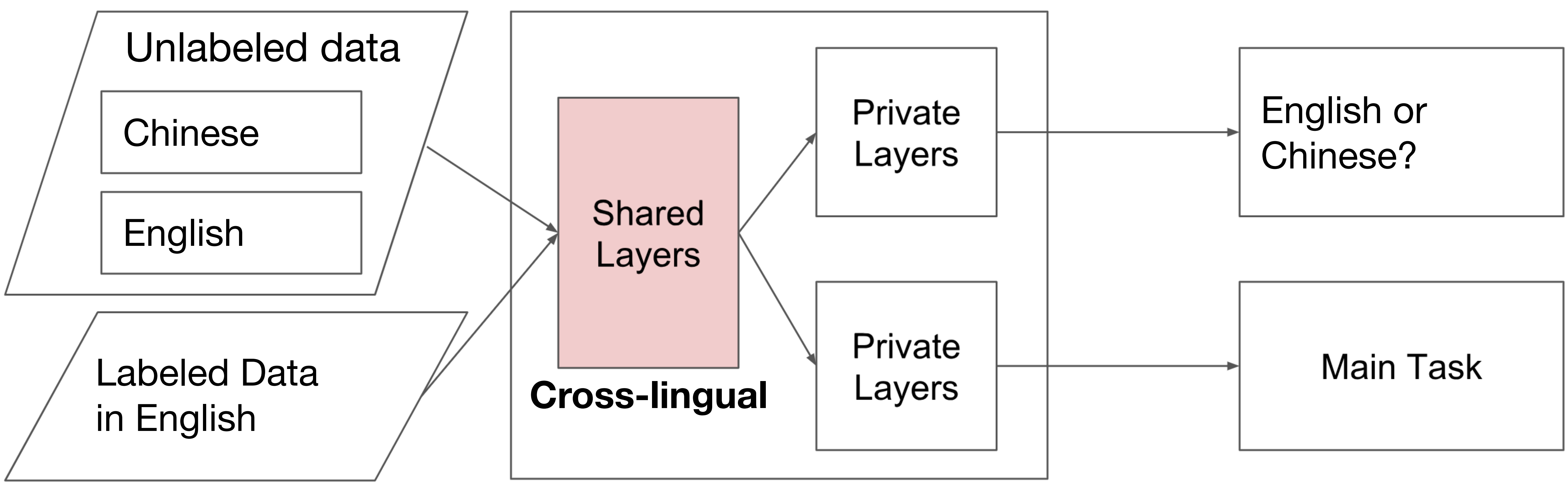
- The feature extractor finds the domain-independent features between the source domain and the target domain.



$$Loss = \lambda * l_{main} - (1-\lambda) * l_{discriminator}$$

Language Discrimination

- Some languages such as English are much more resourceful
- English as source domain, and we aim train the model with labeled data (in English) to deal with the data in another language.



$$Loss = \lambda * l_{main} - (1-\lambda) * l_{discriminator}$$

Cross-lingual Natural Language Inference