De-Identification of free-text

Examiners: Professor Dr. Hannah Bast

Dr. Fang Wei-Kleiner

Albert-Ludwigs-Universität Freiburg

Sameed Hayat

Department of Computer Science
Chair for Algorithms and Data Structure

Problem Statement

Introduction

- Companies and research institution collect and store lot of data
- Legally requirement to protect individual’s privacy
- De-identification means removing identifying information from a dataset so that individual data cannot be linked with specific individual
- **Goals:**

    1. De-identified document should no longer be associated with the person
    2. De-identified record should retain as much information as possible

Motivation

1. Enable new data centric use cases
   - **Medical domain**
     - Obtain insights to assist clinical practice in treatment design or risk assessment, etc.
   - **Corporate domain:**
     - Classification of customer complaints, chatbots, etc.

2. Demonstrate digital responsibility
   - Respect consumer trust and support compliance to regulations such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act of 1996 (HIPAA)

Approach and Technique

De-Identification Task

- 2014 i2b2/UTHealth shared task featured a track focused on the de-identification of longitudinal medical records
- De-identified using the guidelines provided by HIPAA
- A set of 1304 longitudinal medical records describing 296 patients.

De-Identification Task

PHI Category	Sub-Category
NAME	PATIENT, DOCTOR, USERNAME
PROFESSION	None
LOCATION	HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER
AGE	None
DATE	None
CONTACT	PHONE, FAX, EMAIL, URL, IPADDRESS
IDs	SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER

Example: i2b2 2014

ENTITIES: **NAME**, **PROFESSION**, **LOCATION**, **AGE**, **DATE**, **CONTACT**, **IDs**

Record date: **2074-10-01**
Office Note **Bobbie Albert** #**7000963** Tuesday, **October 01, 2074**

Reason for visit

Mr. **Albert** is a **39-year-old** American man status post bilateral lung transplantation due to cystic fibrosis and a history of HCV. He now presents with rising creatinine over the past three months and is referred by Dr. **Benjamin Earnest**.

…

Example: i2b2 2014

ENTITIES: **NAME**, **PROFESSION**, **LOCATION**, **AGE**, **DATE**, **CONTACT**, **IDs**

…

Social-, Work-, and Family- History

…

Married with other lung transplant recipient (tx 9 years ago). No family history of renal disease. Nonsmoker, **Marketing Manager**.

…

Mr. **Albert** has progressive renal insufficiency, he does currently not require renal replacement therapy. Contributors may be Cyclosporine toxicity and intermittent dehydration. However, the recent more rapid rise makes biphosphonate toxicity or HCV-associated renal disease a worthy consideration.

Thomas Yockey, MD.

Challenges

Following can be common sources of ambiguity causing a simpler algorithm to fail:

1. Overlap of words that can be names of people and medical terminologies.

   "Mr. Parkinson" is PHI, while "Parkinson's disease" is not

2. The names may be very uncommon or misspelled.
3. There is no standard data formatting scheme (bulleted data, paragraphs, tabular form etc).

Challenges

4. Fragmented incomplete utterances

5. Domain specific language

Methods

- Automated de-identification systems can be classified into two categories

1. Rule Based Methods
2. Machine Learning based Methods
    1. Feature-engineered supervised systems
    2. Feature-inferring neural network systems

Rule Based Methods

- Typically rely on patterns, expressed as regular expressions, dictionary look-ups, and heuristics, defined and tuned by humans.

PROS

- They do not require any labeled data (aside from labels required for evaluating the system).
- Easy to implement and interpret.

CONS

- Not robust to language changes (e.g., variations in word forms, typographical errors, or infrequently used abbreviations)
- Cannot easily take into account the context (e.g., “Mr. Parkinson” is PHI, while “Parkinson’s disease” is not).

Machine Learning based Methods

Feature-engineered supervised systems

- Supervised machine learning models learn to make predictions by training on example inputs and their expected outputs, and can be used to replace human curated rules.

PROS

- Typically more generalizable than rule-based methods
- Automatically learn complex patterns

CONS

- Require a decent-sized labeled dataset
- Difficult to interpret
- Much feature engineering: quality features are challenging and time-consuming to develop

Machine Learning based Methods

Feature-inferring neural network systems

- Recent approaches to natural language processing based on artificial neural networks (ANNs) do not require handcrafted rules or features.

PROS

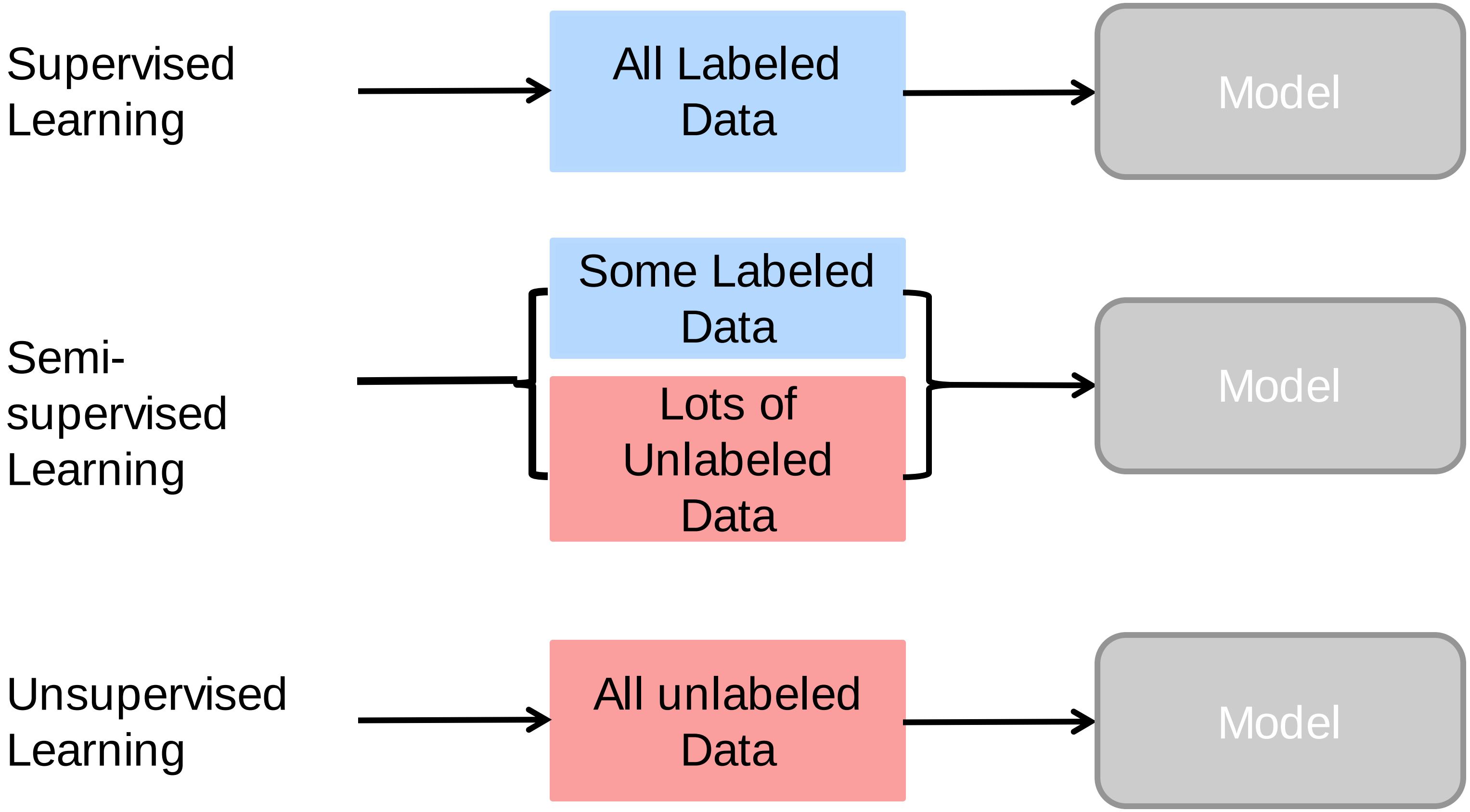
- Robust to language and domain changes
- Low maintenance costs
- Take context into account

CONS

- Require a decent-sized labeled dataset
- Difficult to interpret.

Preliminaries

Types of Machine Learning



Supervised Learning

All Labeled Data

Model

Semi-supervised Learning

Some Labeled Data

Lots of Unlabeled Data

Model

Unsupervised Learning

All unlabeled Data

Model

Embeddings

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation

- **Classical word embeddings**
  - Same embedding vector is generated for a specific word, regardless of the context

- **Contextual word embeddings**
  - Contextual embeddings aim to capture word semantics in different contexts.
  - Capture complex characteristics of word use and how they differ across language contexts

Neural Network for Sequence Labeling

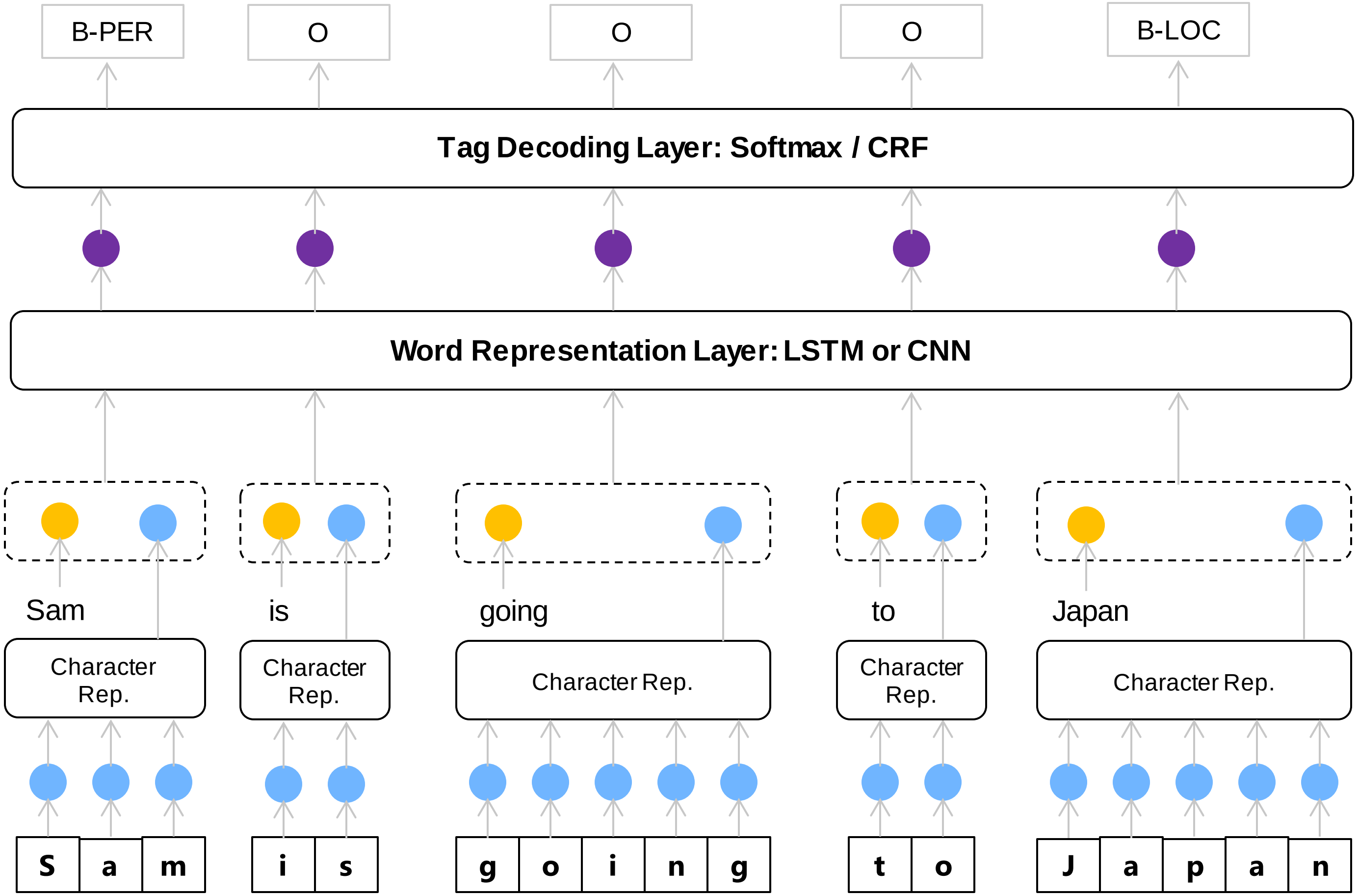
Sequence Labeling

- Assignment of a categorical label to each member of a sequence

- Most of the neural models used for sequence labelling have three main components, which are as follows.

- 1. Character representation layer
- 2. Word representation layer
- 3. Tag decoding layer

Neural Network for Sequence Labeling



Neural Network for Sequence Labeling

Character representation layer:

- Two ways to extract character representation

1. CNN Character Representation Layer
2. LSTM Character Representation Layer

Neural Network for Sequence Labeling

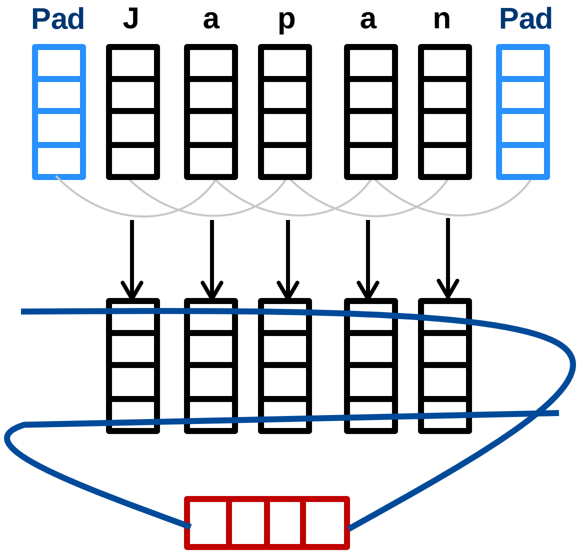
CNN Character Representation Layer

Character Embedding

Convolution

Max Pooling

Character Representation



Neural Network for Sequence Labeling

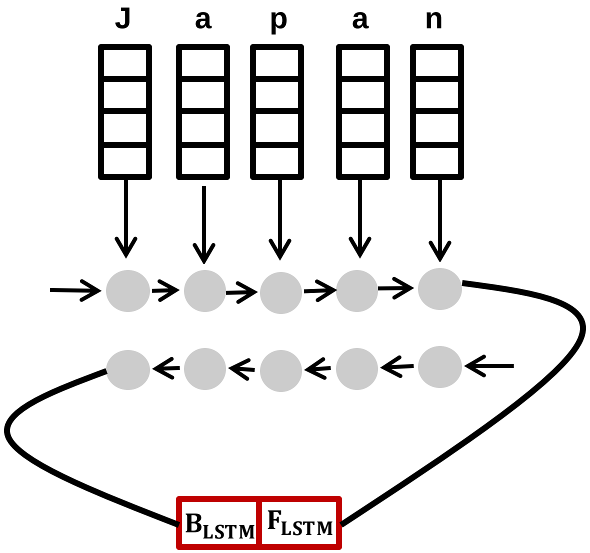
LSTM Character Representation Layer

Character Embedding

Forward LSTM

Backward LSTM

Character Representation



Approach

- In this thesis, we tend to investigate two semi-supervised approaches which are as follows

1. Character-level contextual embedding (Flair)

2. Cross-View Training

Flair Embedding

Character language model to produce a novel type of word embedding which we refer to as contextual string embeddings

Obtaining Flair embedding could be divided into two steps.

1. Training Language Model
2. Extracting Word Representations

Flair Embedding

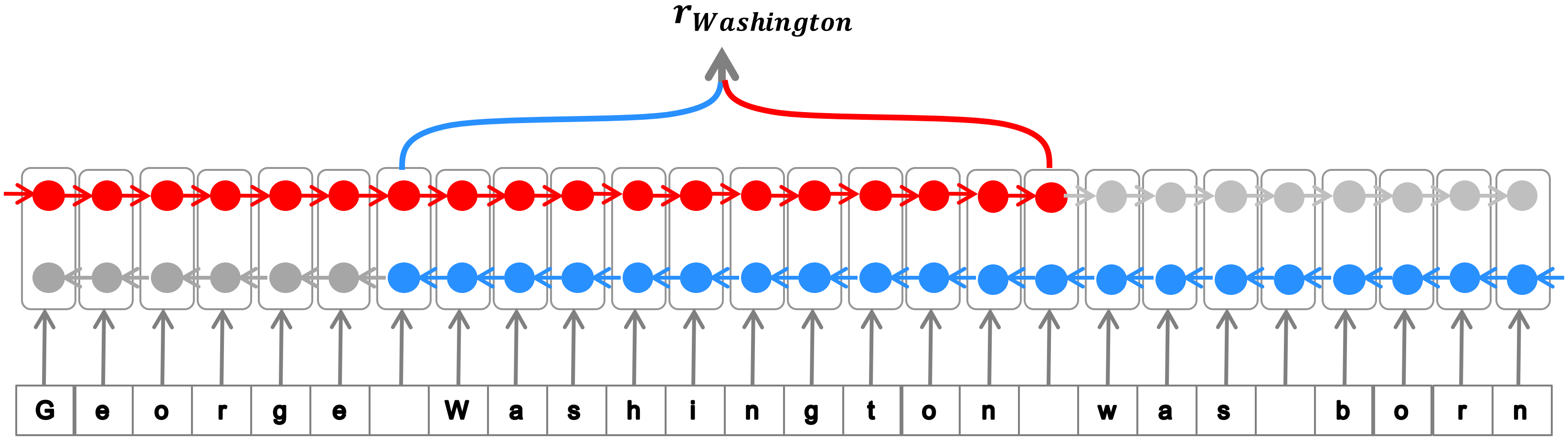
Training Language Model

- Sequence of characters passed to LSTM
- At each point in the sequence model is trained to predict the next character in the sequence
- Goal to estimate a good distribution over sequences of characters reflecting natural language production

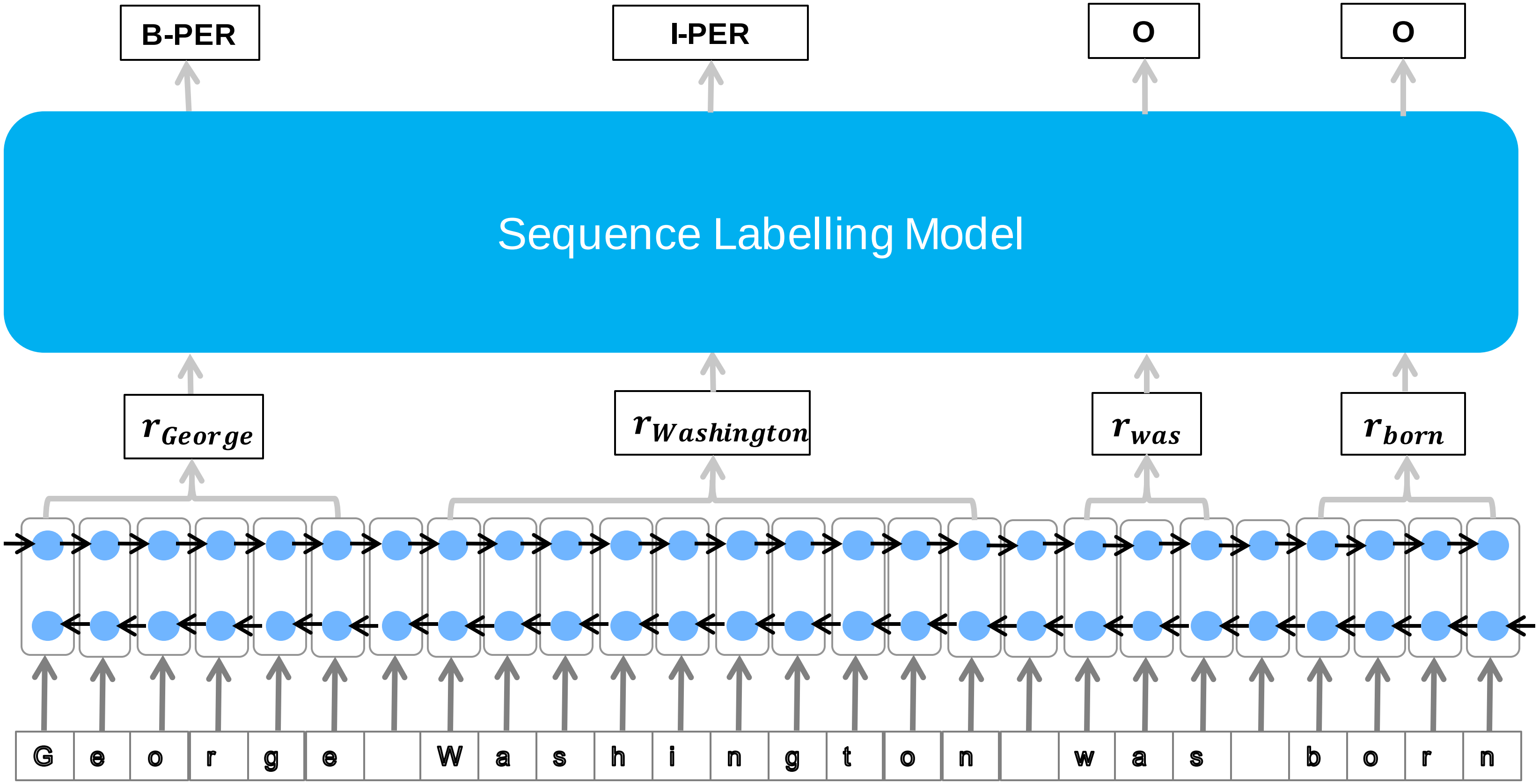
Flair Embedding

Extracting Word Representations

- From the **forward language model**, we extract the output hidden states **after the last character** in the word
- From the **backward language model**, we extract the output hidden states **before the first character** in the word



Flair Embedding for Sequence Labelling



Pros and Cons of Flair

Advantages

- Different embeddings for polysemous words
- Handles rare and misspelled words
- Handles subword structures such as prefixes and endings

Disadvantages

- Learn generally useful representations
- Large model size

Cross View Training (CVT)

- CVT is an effective training mechanism to make use of labeled and unlabeled data for training the model

- Based on self-learning in a neural world

- Learn representations targeted to a particular task

Cross View Training (CVT)

Classic Self-Training

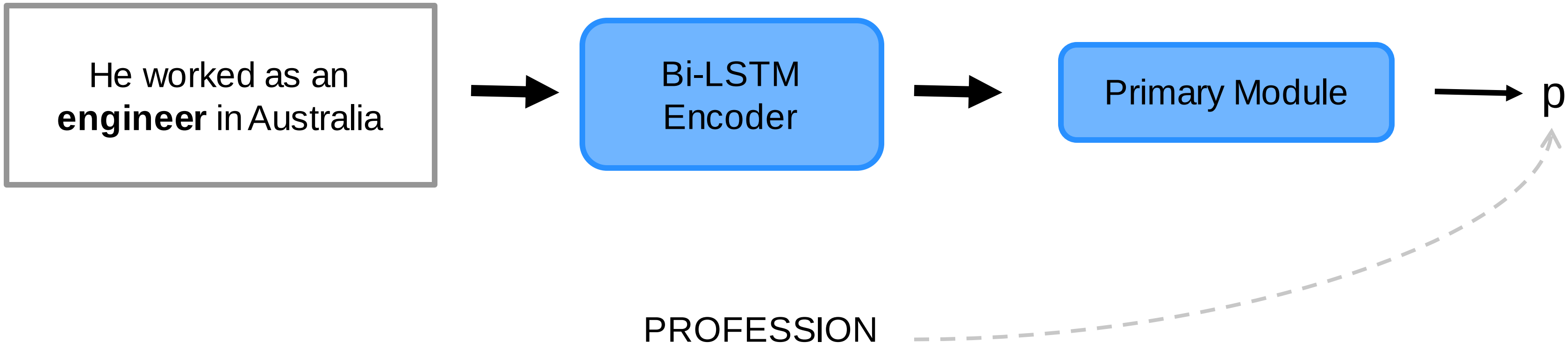
- Alternate learning on labeled and unlabeled examples

- **For Labeled data**
  - Standard supervised learning as in the case of sequence labelling

- **For unlabeled data**
  - Approach is to self label

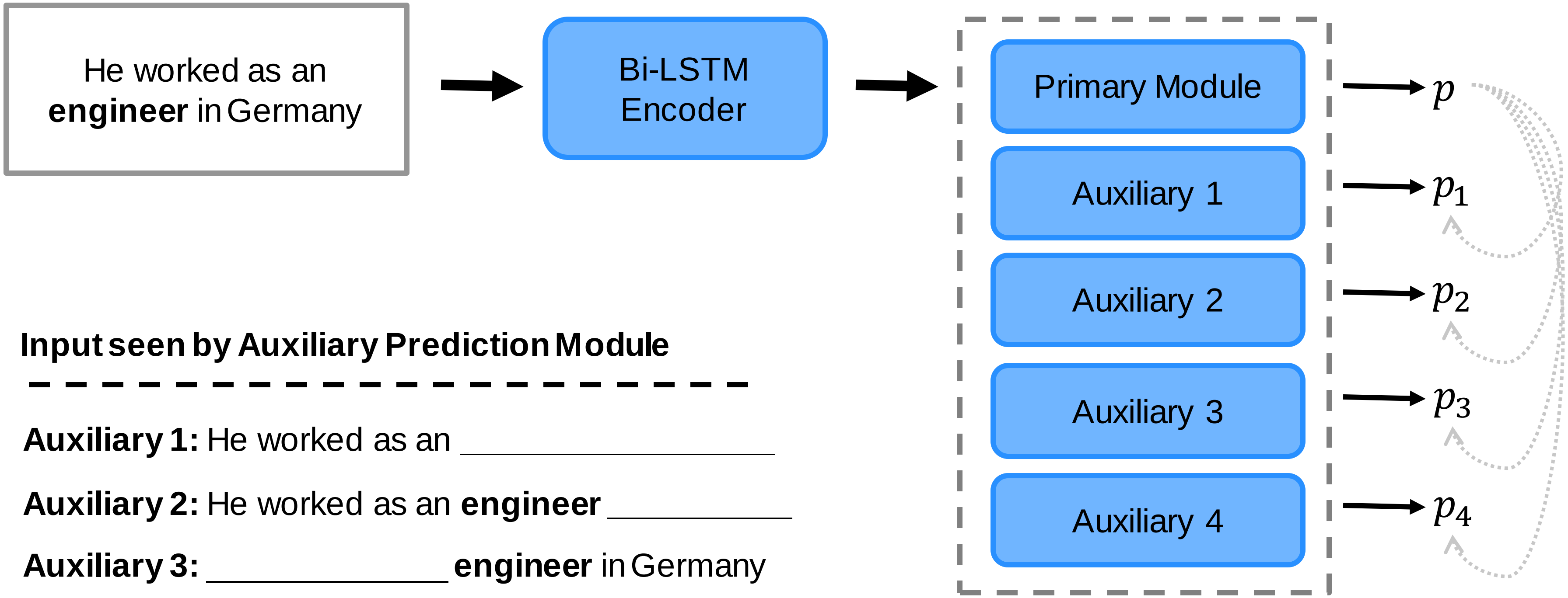
Cross View Training (CVT)

Labeled Data



Cross View Training (CVT)

Unlabeled Data



He worked as an **engineer** in Germany

Bi-LSTM Encoder

Primary Module

Auxiliary 1

Auxiliary 2

Auxiliary 3

Auxiliary 4

Input seen by Auxiliary Prediction Module

Auxiliary 1: He worked as an ____________________

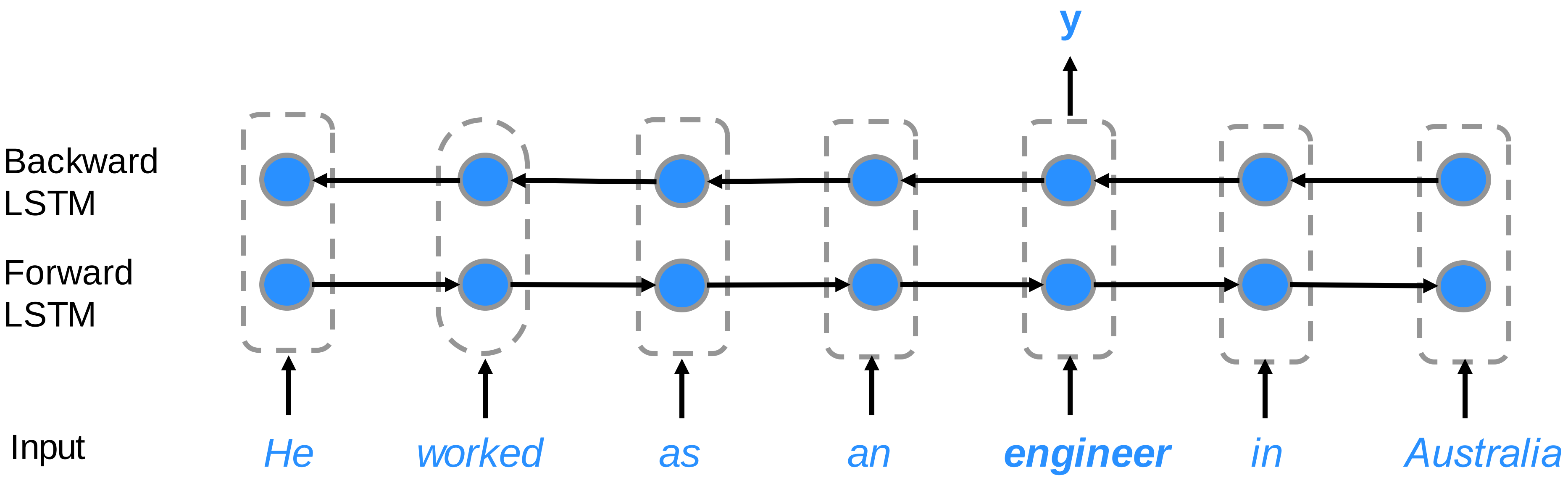
Auxiliary 2: He worked as an **engineer** ___________

Auxiliary 3: _______________ **engineer** in Germany

Auxiliary 4: ________________________ in Germany

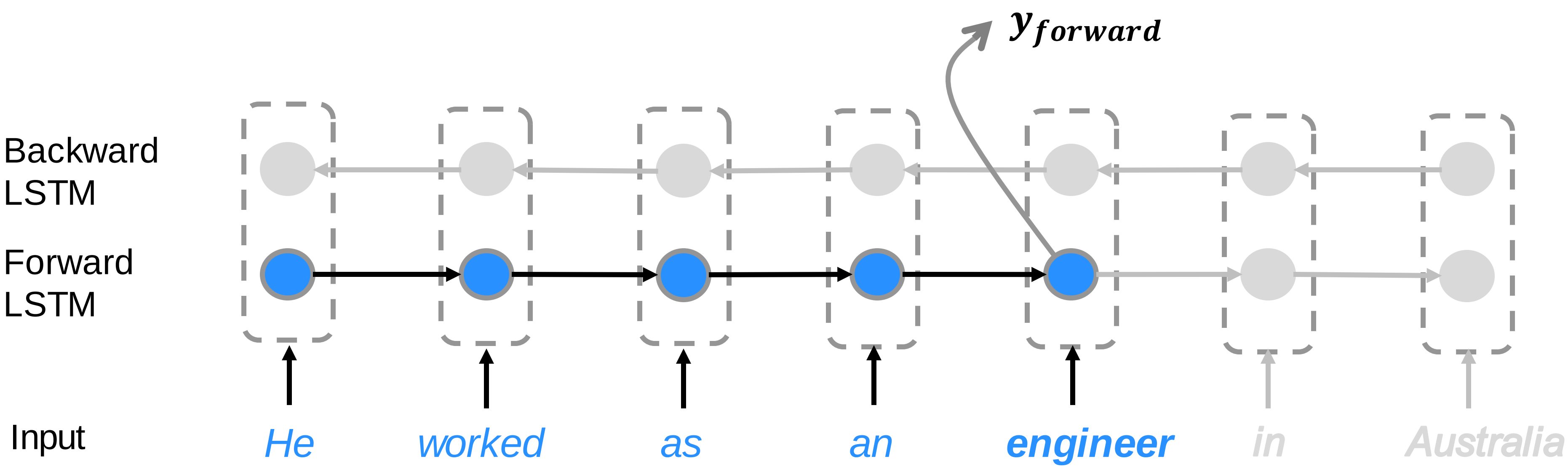
Cross View Training (CVT)

Unlabeled Data: Main Predictor



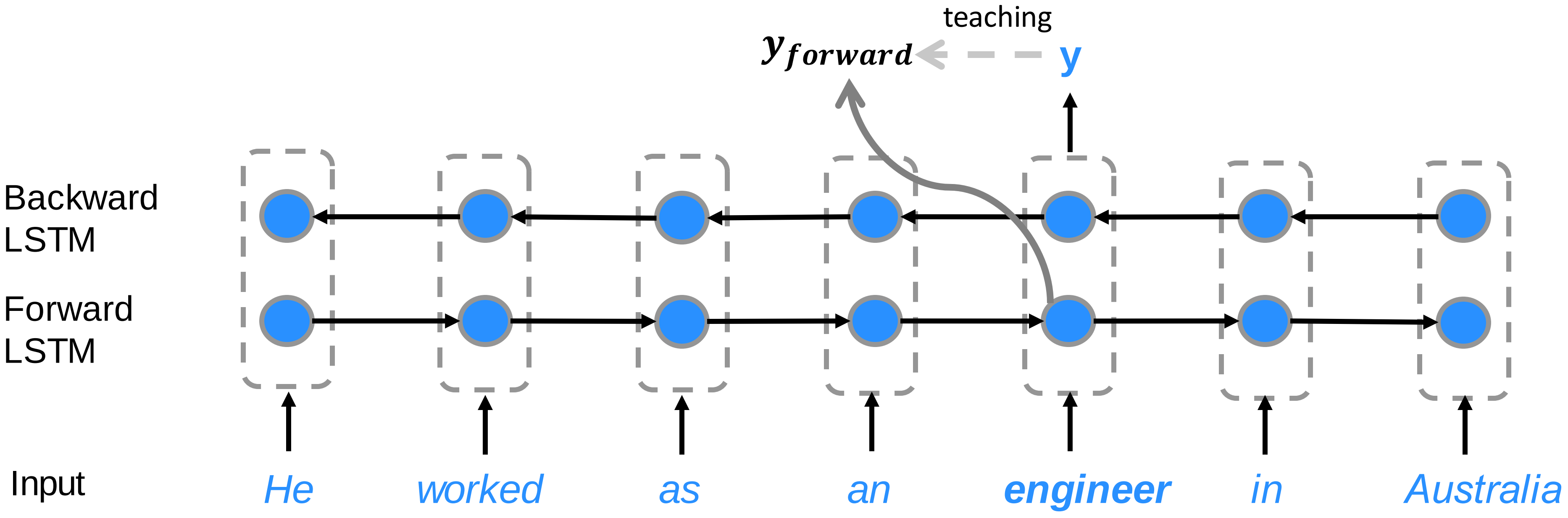
Cross View Training (CVT)

Unlabeled Data: Forward Predictor

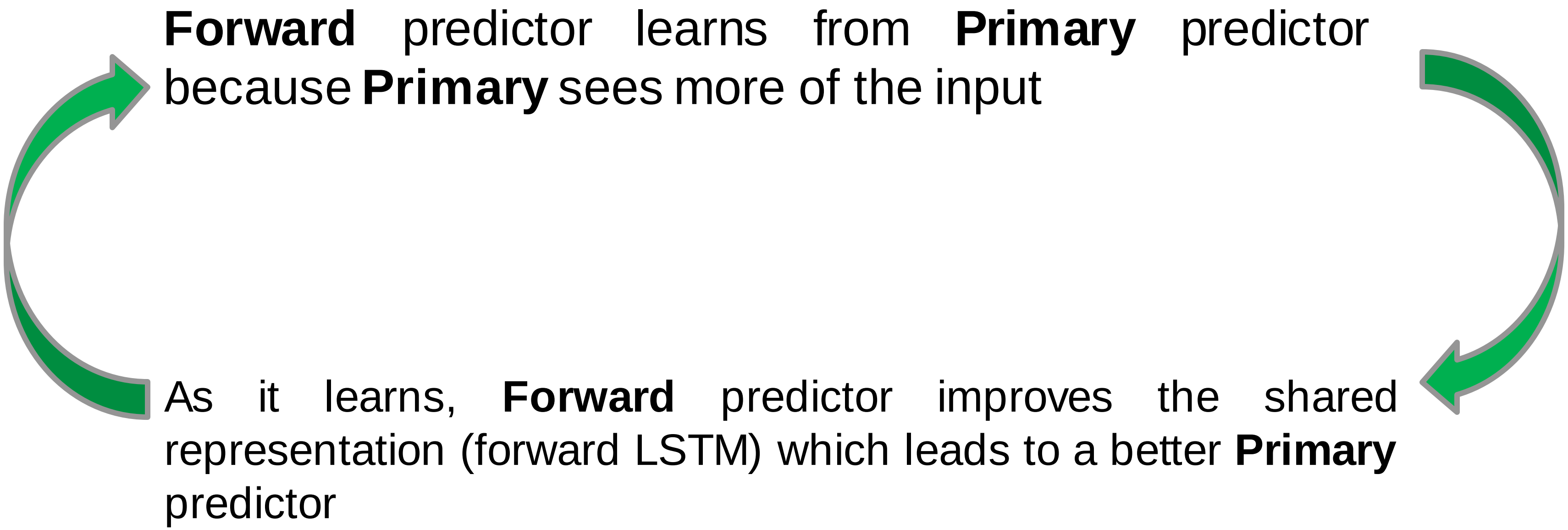


Cross View Training (CVT)

Unlabeled Data: Forward Predictor



Cross View Training (CVT)

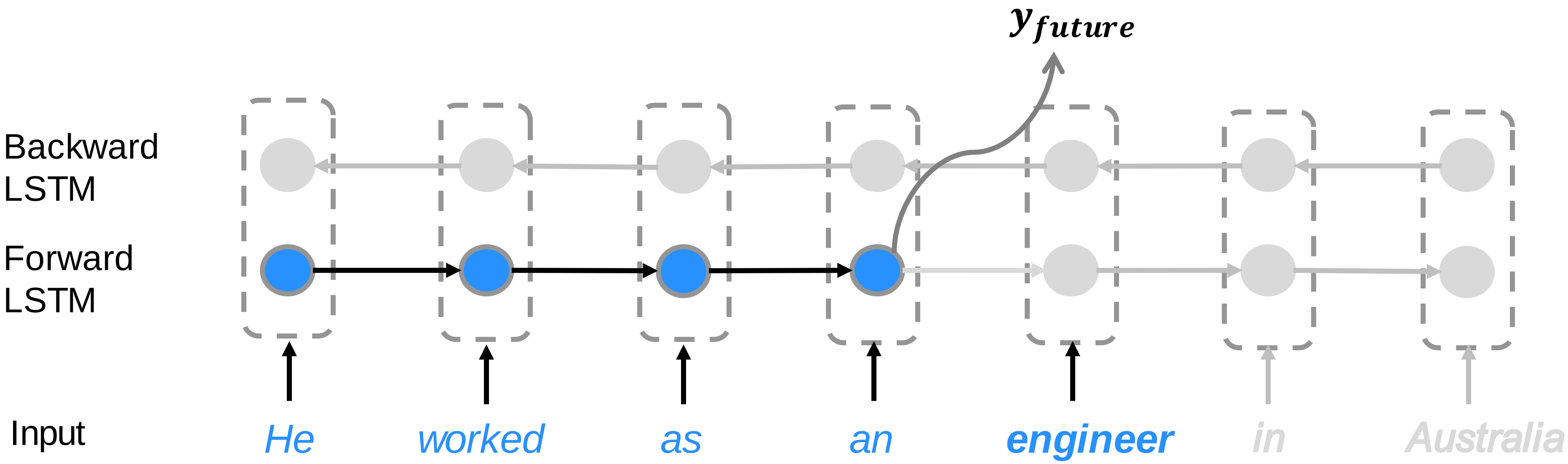


Forward predictor learns from **Primary** predictor because **Primary** sees more of the input

As it learns, **Forward** predictor improves the shared representation (forward LSTM) which leads to a better **Primary** predictor

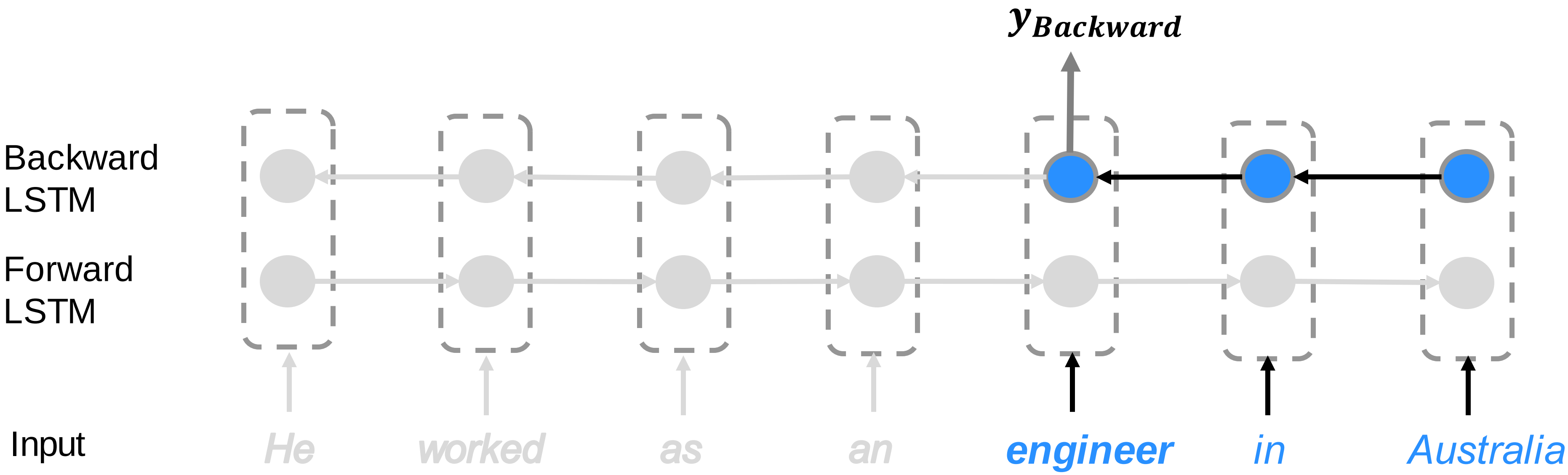
Cross View Training (CVT)

Unlabeled Data: Future Predictor



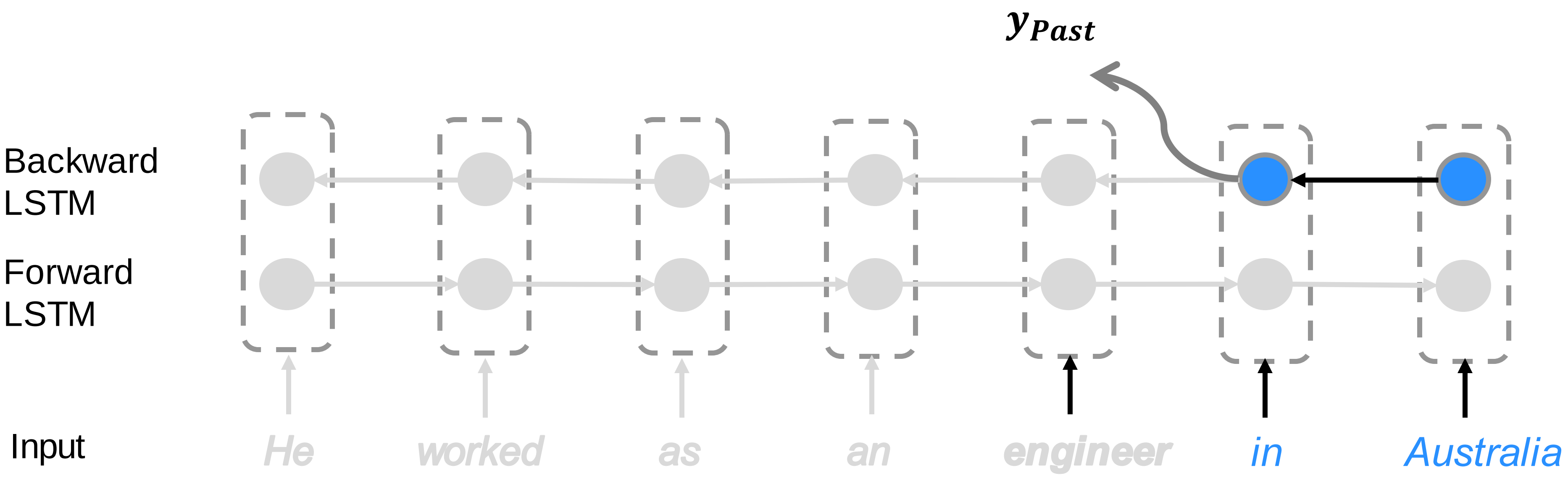
Cross View Training (CVT)

Unlabeled Data: Backward Predictor



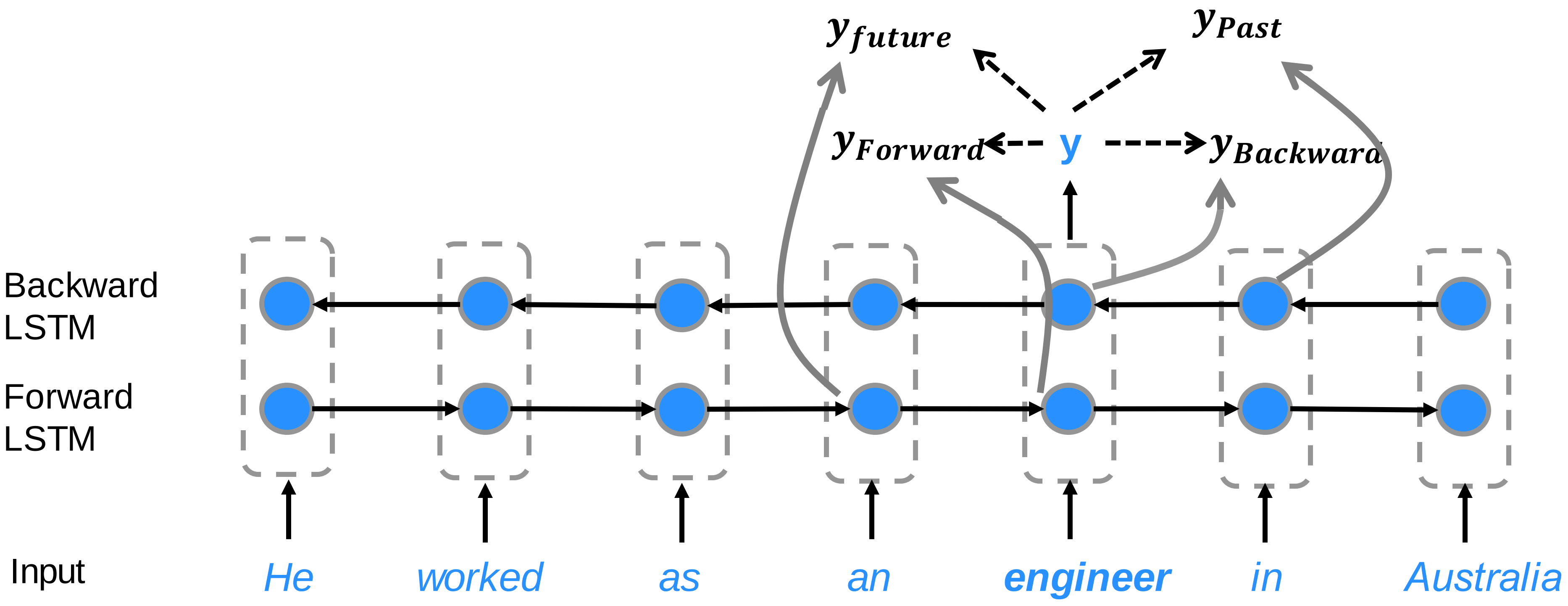
Cross View Training (CVT)

Unlabeled Data: Backward Predictor



Cross View Training (CVT)

Unlabeled Data: Past Predictor



Pros and Cons of CVT

Advantages

- Effective usage of unlabeled data as model learn representations targeted to a particular task
- Comparable or better accuracy
- Works well for small labeled datasets

Disadvantages

- Requires in-domain unlabeled data
- Have to train for each task

Evaluation: Setup and Main Results

Datasets for Unsupervised Learning

Two datasets for unsupervised learning task

- **1 Billion Word Language Model Benchmark**
  - Dataset based on WMT 2011 News Crawl data

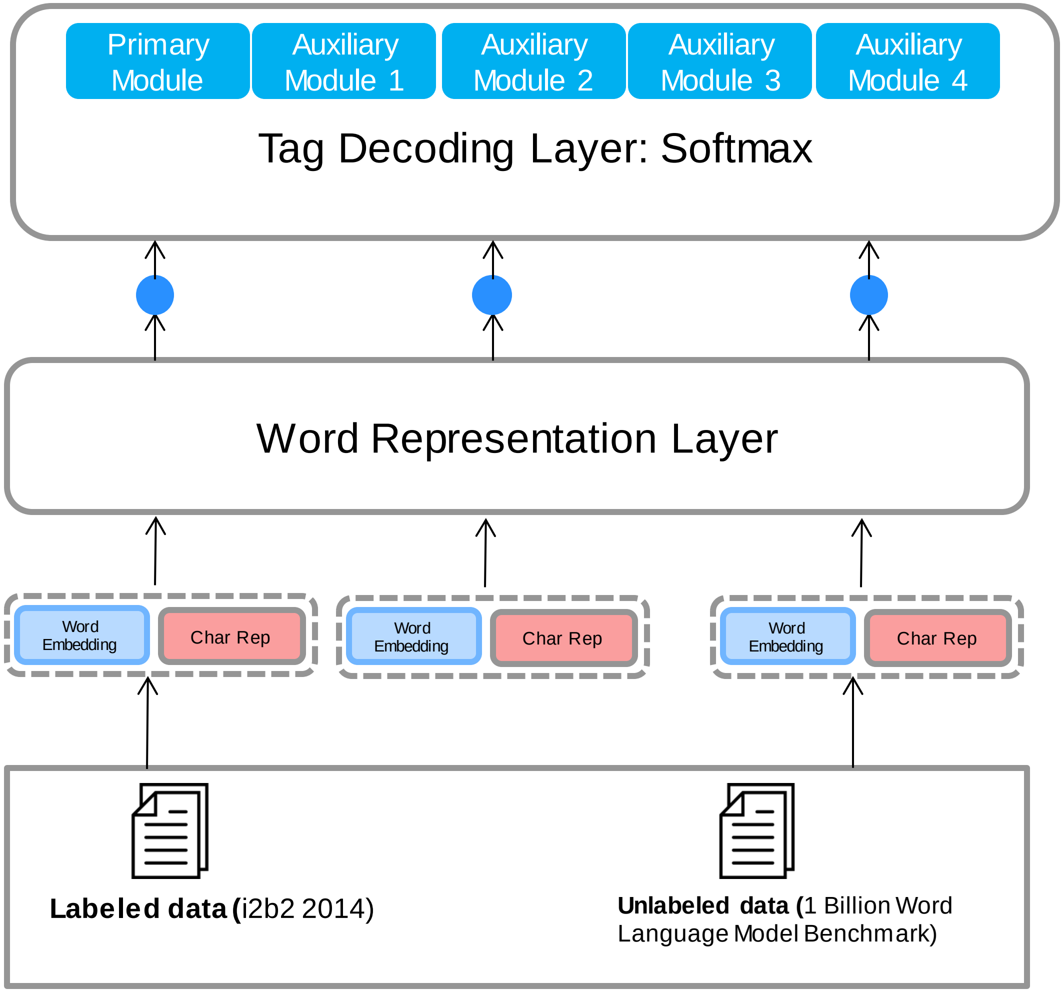
- **MIMIC-III**
  - Containing information related to patients admitted to critical care units
  - Discharge summaries, which comprises of 59652 notes

Experiments and Results

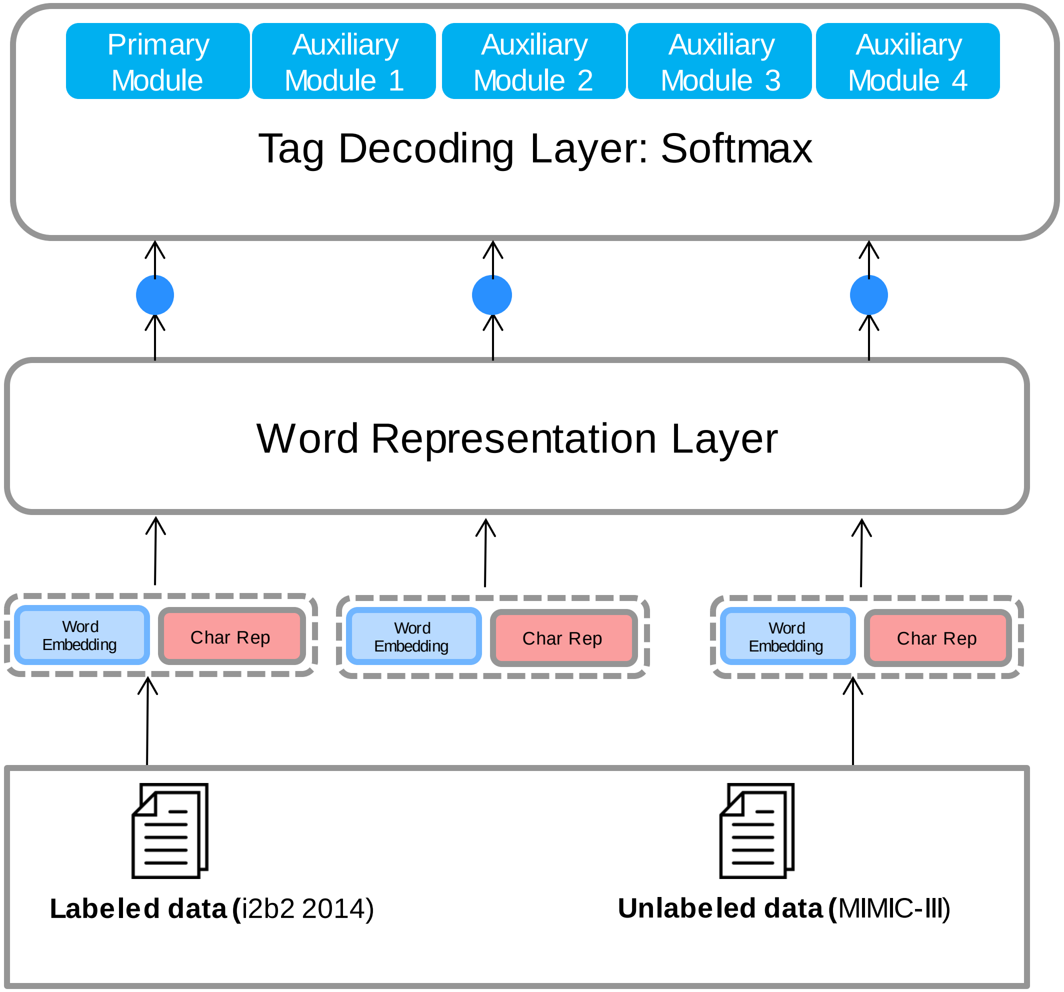
We examine two models with different configurations, which are as follows

- CVT(1b)
- CVT(MIMIC)
- Flair(1b)
- Flair(MIMIC)

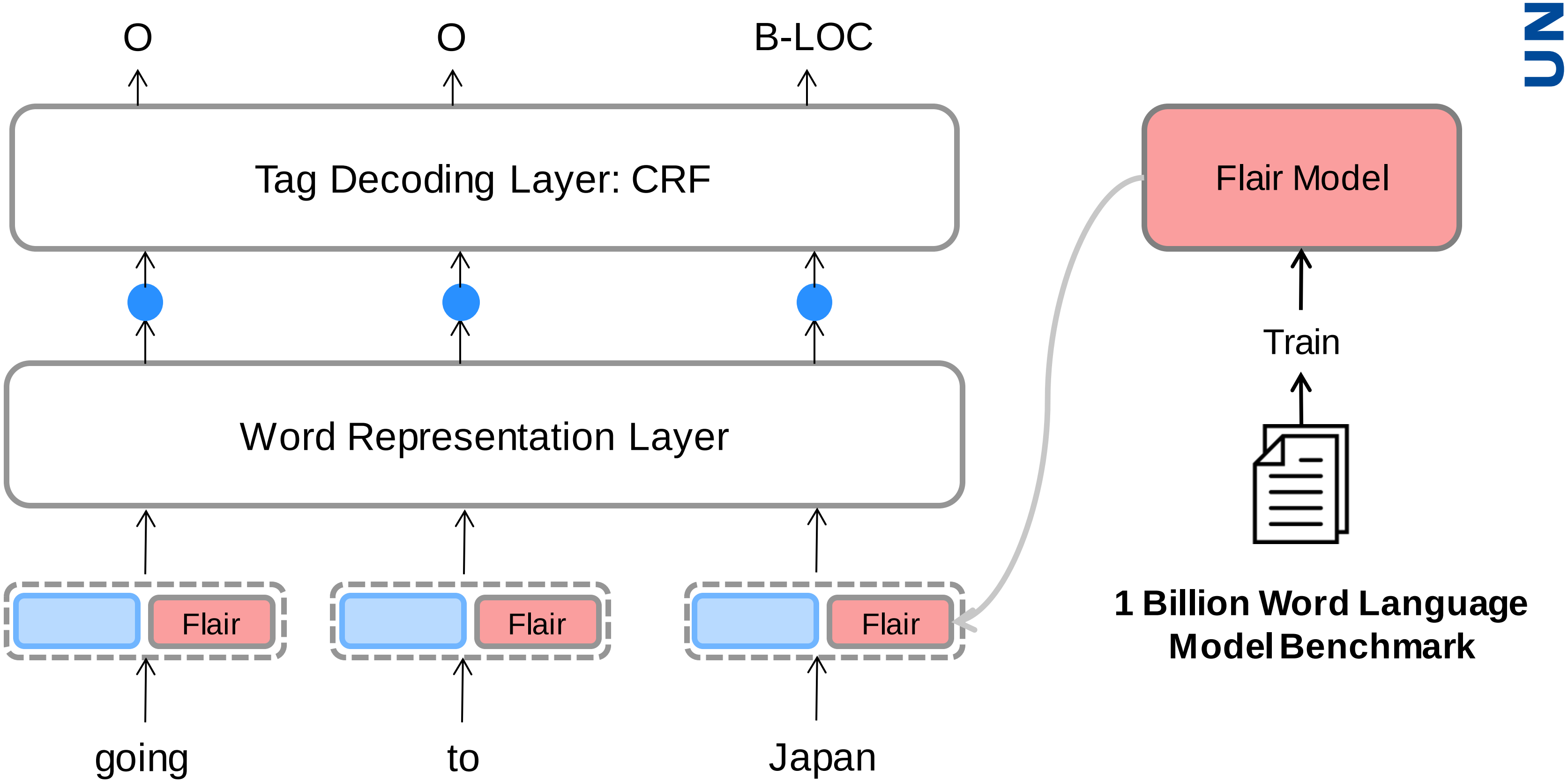
CVT(1b)


Primary Module
Auxiliary Module 1
Auxiliary Module 2
Auxiliary Module 3
Auxiliary Module 4
Tag Decoding Layer: Softmax
Word Representation Layer
Word Embedding
Char Rep
Word Embedding
Char Rep
Word Embedding
Char Rep
Labeled data (i2b2 2014)
Unlabeled data (1 Billion Word Language Model Benchmark)

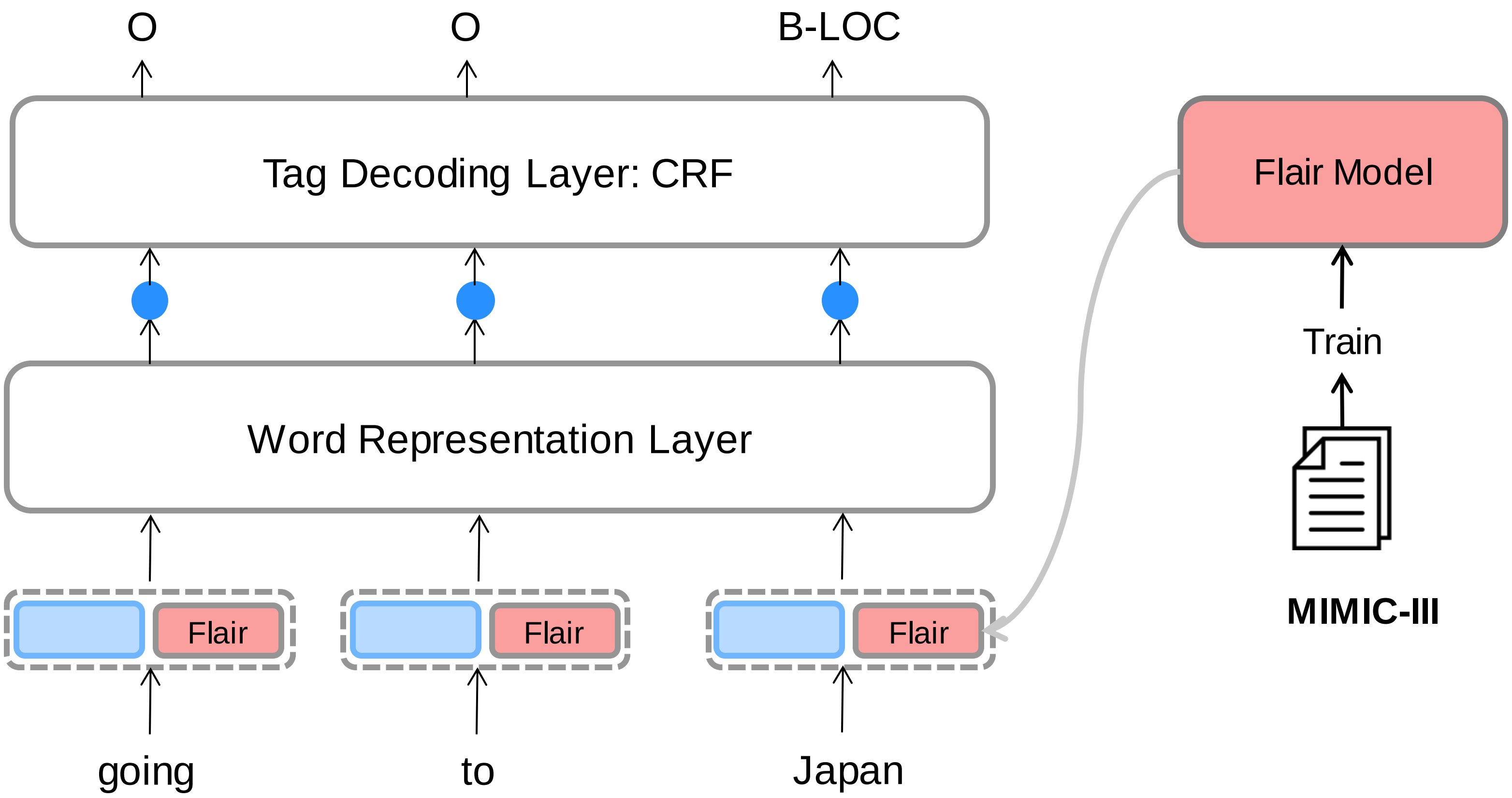
CVT(MIMIC)


Primary Module
Auxiliary Module 1
Auxiliary Module 2
Auxiliary Module 3
Auxiliary Module 4
Tag Decoding Layer: Softmax
Word Representation Layer
Word Embedding
Char Rep
Word Embedding
Char Rep
Word Embedding
Char Rep
Labeled data (i2b2 2014)
Unlabeled data (MIMIC-III)

Flair(1b)


O
O
B-LOC
Tag Decoding Layer: CRF
Word Representation Layer
Flair
Flair
Flair
going
to
Japan
Flair Model
Train
1 Billion Word Language
Model Benchmark

Flair(MIMIC)

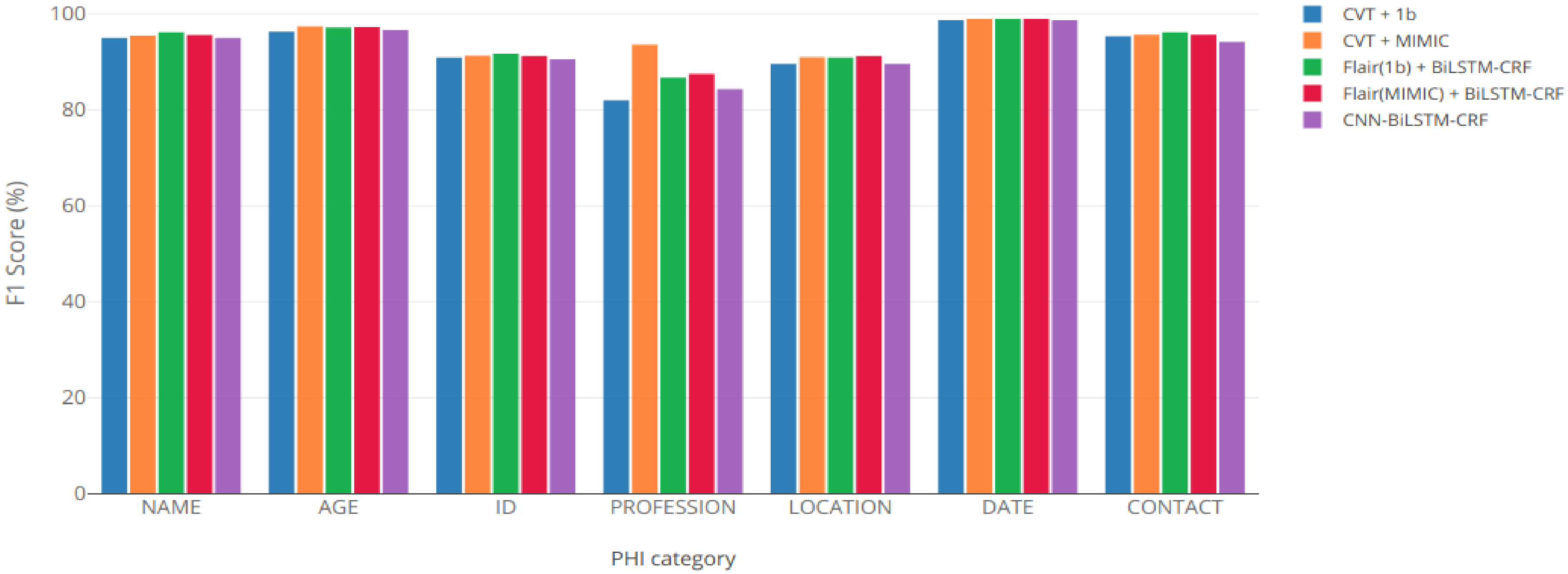

O
O
B-LOC
Tag Decoding Layer: CRF
Flair Model
Train
Word Representation Layer
MIMIC-III
Flair
Flair
Flair
going
to
Japan

Results

Model	Precision	Recall	F1-Score
ANN + CRF [Dernoncourt et al. 2017]	97.92	**97.83**	97.87
Elmo + BiLSTM-CRF [Khin et all.. 2018]	98.30	97.37	97.83
BiLSTM-CRF	98.03	97.20	97.61
CVT(1b)	97.96	97.27	97.62
CVT(MIMIC)	98.22	97.69	97.95
Flair(1b)	**98.46**	97.52	**97.99**
Flair(MIMIC)	98.28	97.61	97.94

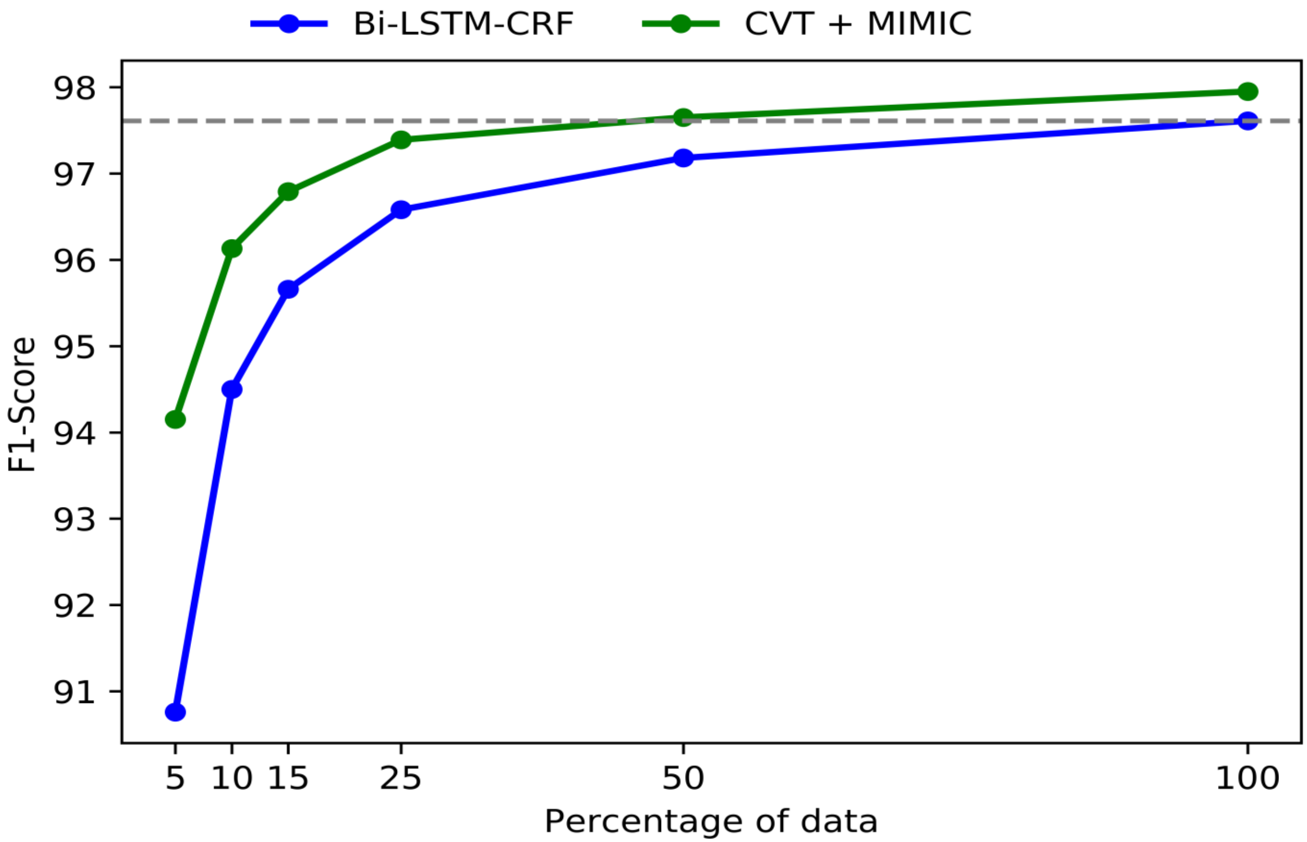
F1-Score (%) on HIPAA-PHI categories on 2014 i2b2 / UTHealth shared task Track 1

Results



F1-Score (%) on HIPAA-PHI categories on 2014 i2b2 / UTHealth shared task Track 1

Training Models on Small Dataset



F1-Score (%) on HIPAA-PHI categories on 2014 i2b2 / UTHealth shared task Track 1

Conclusion

- Semi-supervised approaches are effective for the task of sequence modelling in medical domain
- Cross-View Training performs better than purely supervised methods
- Cross-View Training is only effective when in-domain data is available
- Cross-View training achieve same F1-Score as purely supervised using 50% of the data
- Character-level contextual embeddings produce best performance in terms of F1 Score.
- Character-level embeddings works with both in-domain or out-of-domain data

Future Work

- Combine both approaches for de-identification task
- More in-domain unlabeled data
- Multitask learning
- Hyperparameter optimization could be investigated

References

- Clark, K., Luong, M. T., Manning, C. D., & Le, Q. V. (2018). Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

- Akbik, A., Blythe, D., & Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638-1649).

- Khin, K., Burckhardt, P., & Padman, R. (2018). A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. *arXiv preprint arXiv:1810.01570*.

Thank you!

Examples of features used in the CRF model

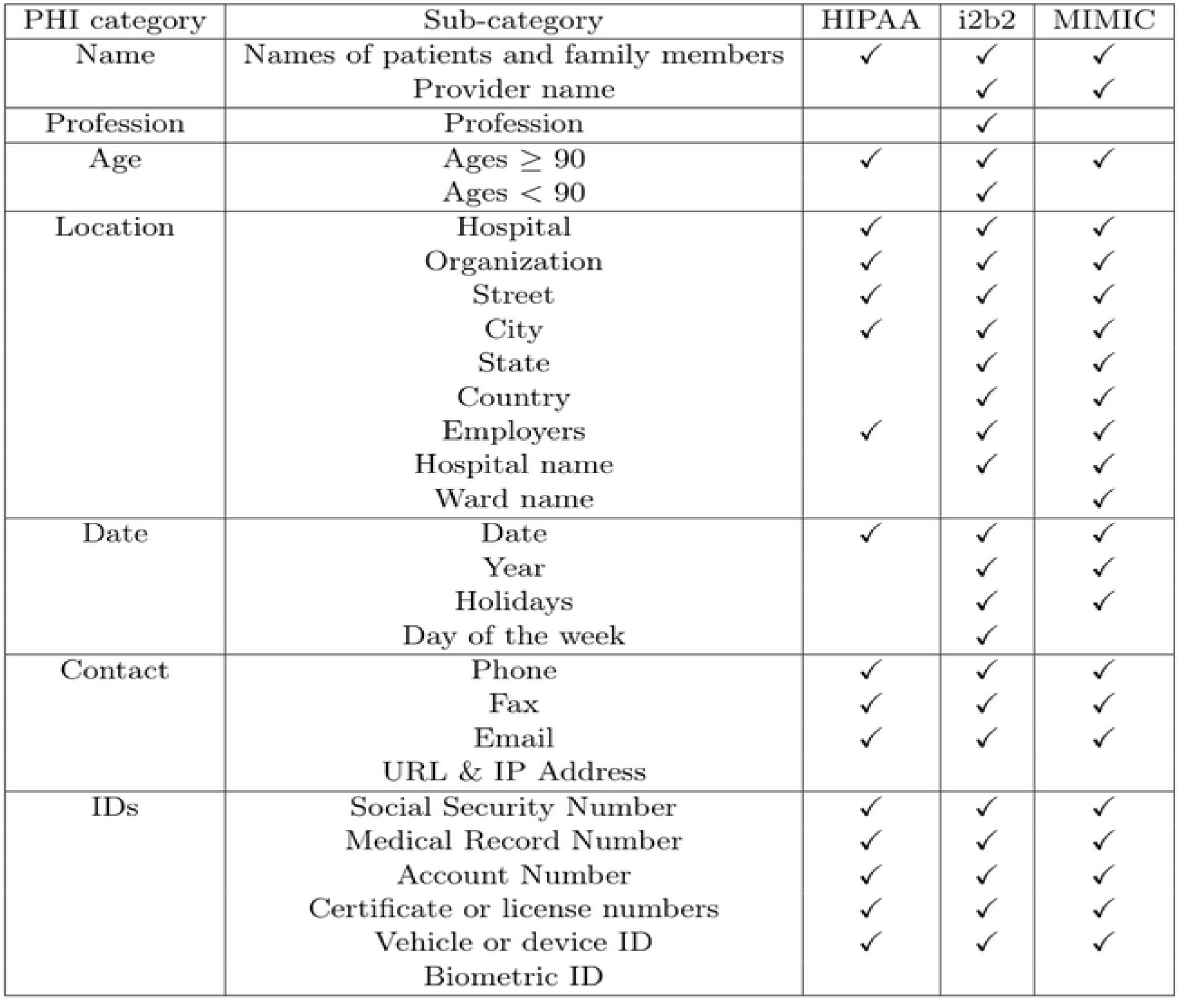
Feature types	Features
Lexical/syntactic	Token, lemma, tense, parts of speech
Morphological	Ends with s, contains a digit, is numeric, is alphabetic, is alphanumeric, is title case, is all lowercase, prefix, suffix
Temporal	Season, month, weekday, time of day
Semantic/wordnet	Hypernyms, senses, lemma names
Gazetteers	First names, last names, medical titles, medical specialties, cities, states (including abbreviations), countries, organizations, professions, holidays
Regular expressions	E-mail, age, date, phone, zip code, ID number, medical record number

Tokenization

- Text is tokenizated aggressively
  - Splitting after all punctuation marks
  - Splitting if number is followed by text like 20yo with “20“ and “yo“

- Split using spaCy heuristics for English with additional rules
  - Split after three newline characters
  - Split for bulleted or numbered list items
  - Split after three dashes

PHI types as defined by HIPAA, i2b2, and MIMIC.



PHI category	Sub-category	HIPAA	i2b2	MIMIC
Name	Names of patients and family members	✓	✓	✓
	Provider name		✓	✓
Profession	Profession		✓	
Age	Ages $\geq$ 90	✓	✓	✓
	Ages $<$ 90		✓	
Location	Hospital	✓	✓	✓
	Organization	✓	✓	✓
	Street	✓	✓	✓
	City	✓	✓	✓
	State		✓	✓
	Country		✓	✓
	Employers	✓	✓	✓
	Hospital name		✓	✓
	Ward name			✓
Date	Date	✓	✓	✓
	Year		✓	✓
	Holidays		✓	✓
	Day of the week		✓	
Contact	Phone	✓	✓	✓
	Fax	✓	✓	✓
	Email	✓	✓	✓
	URL & IP Address			
IDs	Social Security Number	✓	✓	✓
	Medical Record Number	✓	✓	✓
	Account Number	✓	✓	✓
	Certificate or license numbers	✓	✓	✓
	Vehicle or device ID	✓	✓	✓
	Biometric ID			

Tagging Scheme

Smith	is	going	to	San	Francisco
I-PER	O	O	O	B-LOC	I-LOC

IOB Tagging Scheme

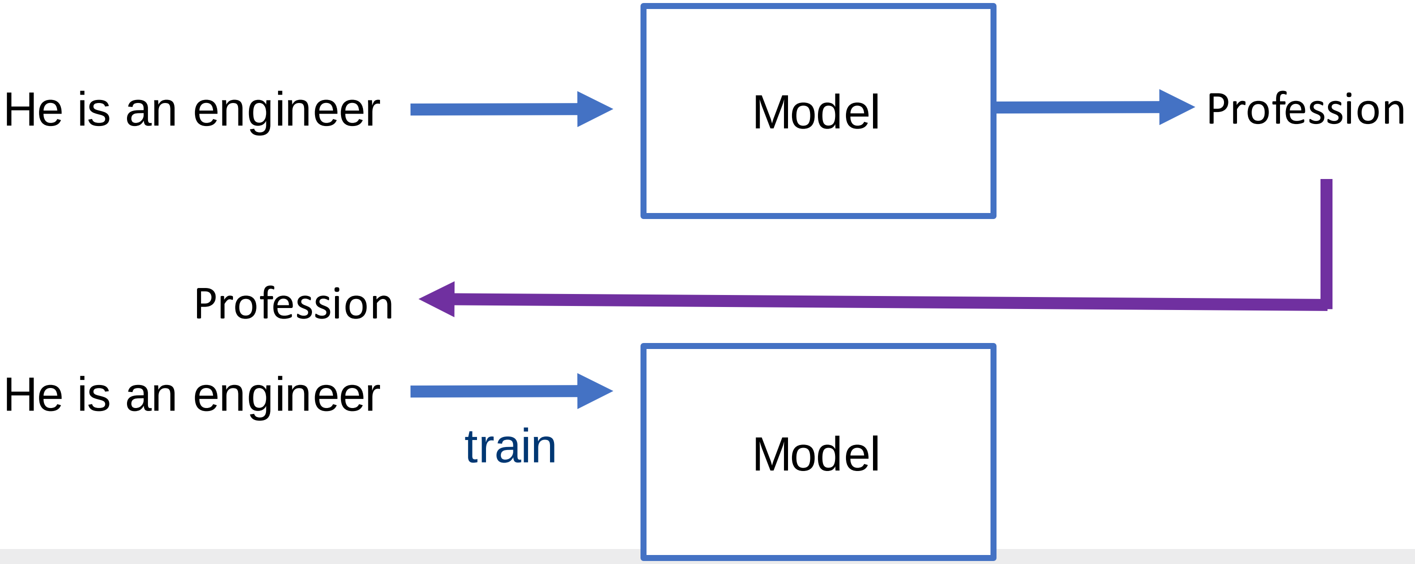
Smith	is	going	to	San	Francisco
B-PER	O	O	O	B-LOC	I-LOC

IOB2 Tagging Scheme

Cross View Training (CVT)

Classic Self-Training

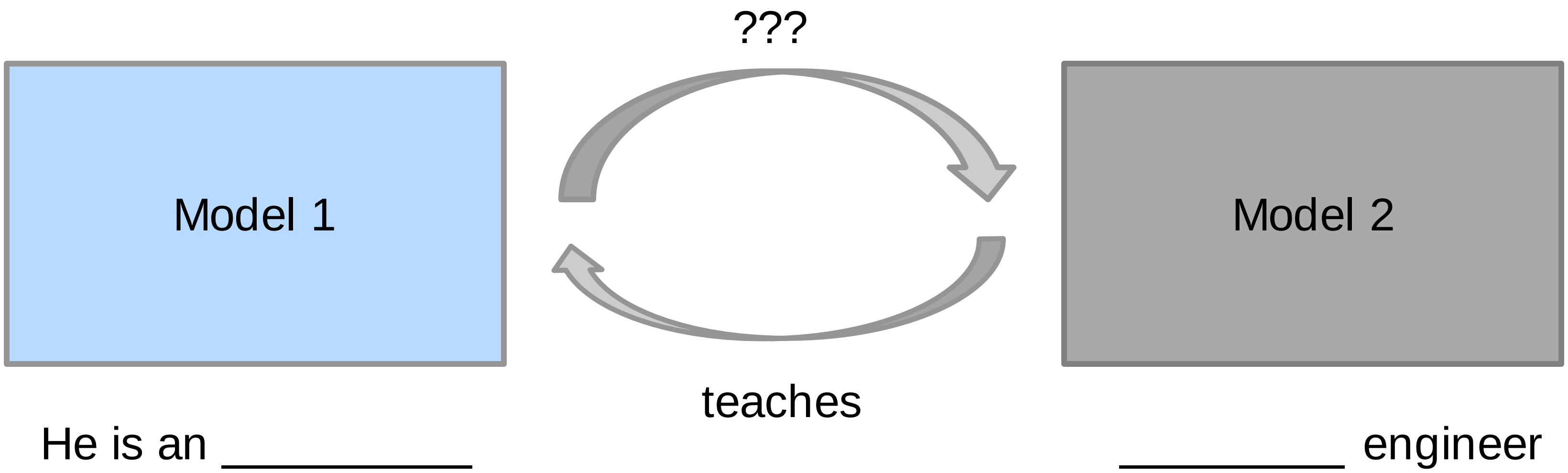
- Alternate learning on labeled and unlabeled examples
- For Labeled data
  - Standard supervised learning as in the case of Named Entity Recognition NER
- For unlabeled data
  - Approach is to self label



Cross View Training (CVT)

Co-Training

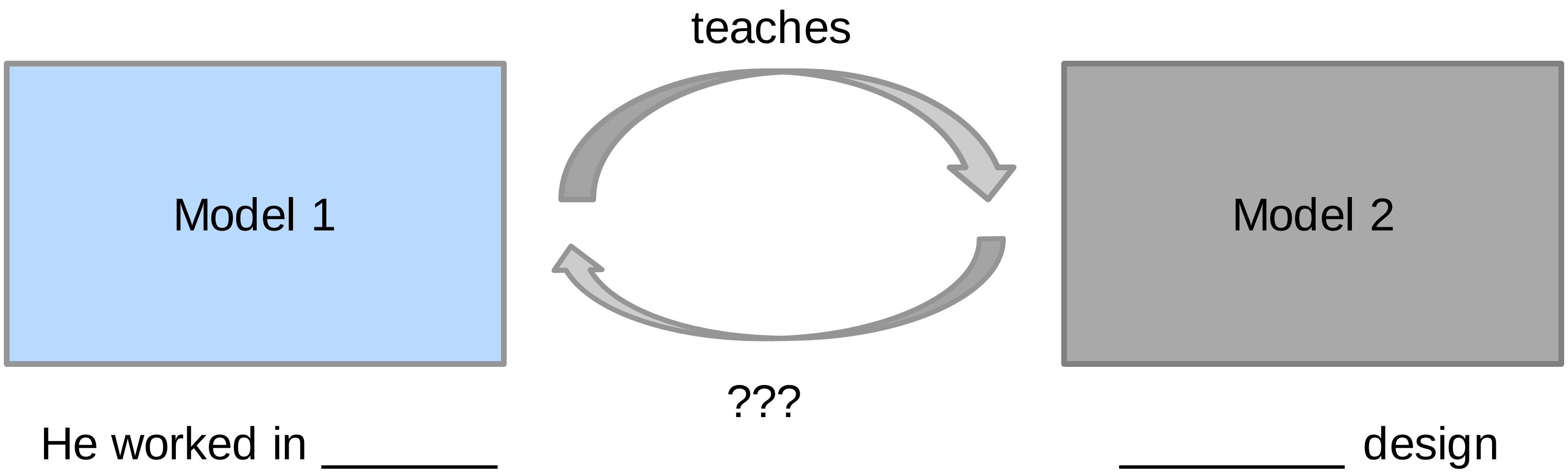
Example: ***He is an engineer***



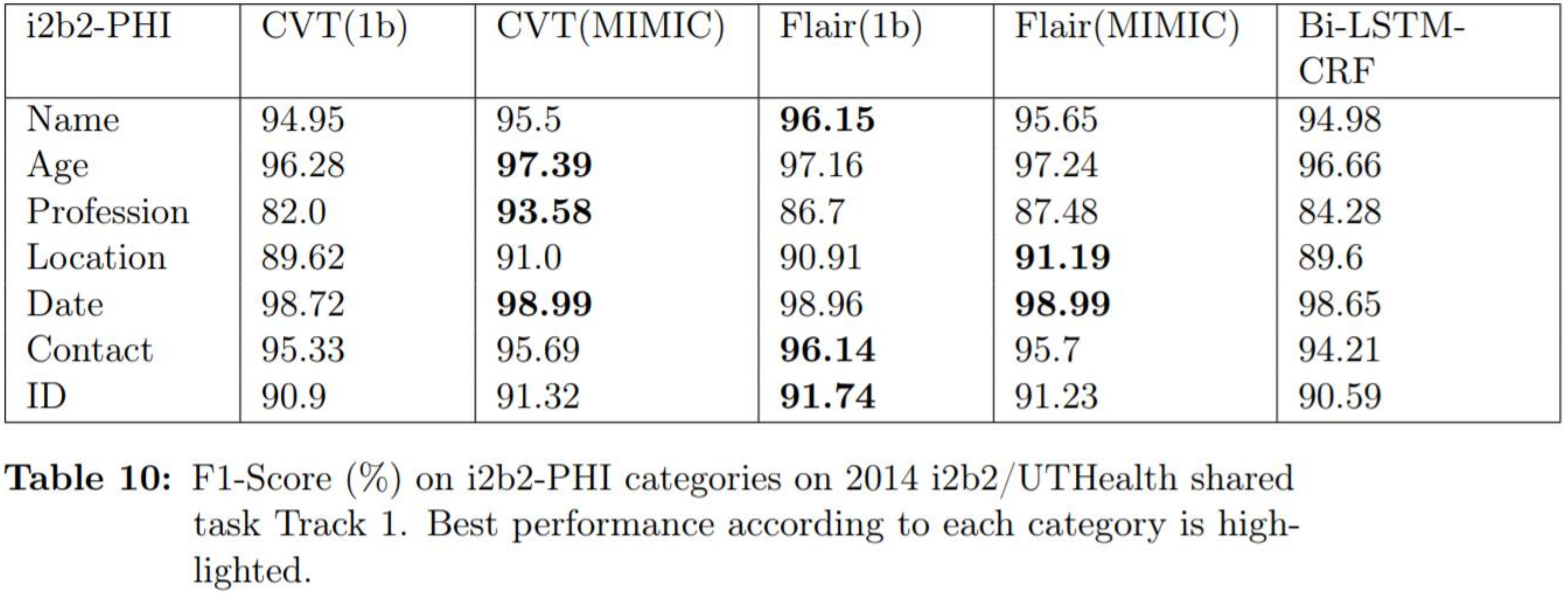
Cross View Training (CVT)

Co-Training

Example: ***He worked in design***



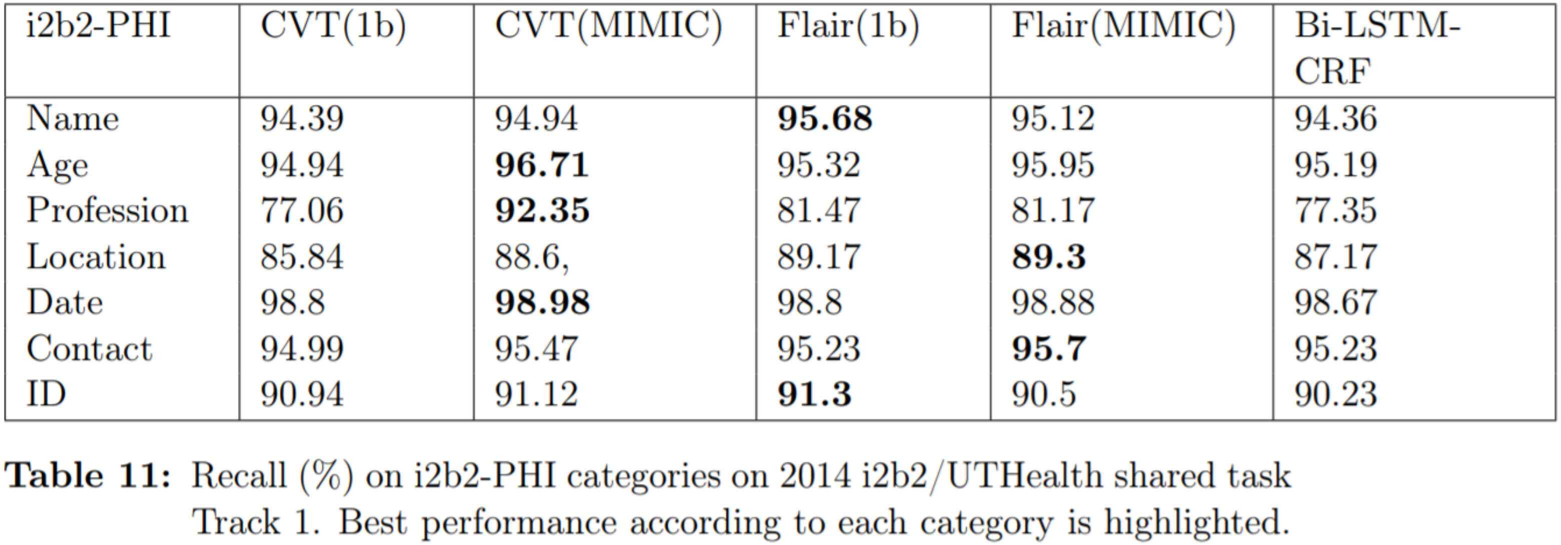
Results



i2b2-PHI	CVT(1b)	CVT(MIMIC)	Flair(1b)	Flair(MIMIC)	Bi-LSTM-CRF
Name	94.95	95.5	**96.15**	95.65	94.98
Age	96.28	**97.39**	97.16	97.24	96.66
Profession	82.0	**93.58**	86.7	87.48	84.28
Location	89.62	91.0	90.91	**91.19**	89.6
Date	98.72	**98.99**	98.96	**98.99**	98.65
Contact	95.33	95.69	**96.14**	95.7	94.21
ID	90.9	91.32	**91.74**	91.23	90.59

Table 10: F1-Score (%) on i2b2-PHI categories on 2014 i2b2/UTHealth shared task Track 1. Best performance according to each category is highlighted.

Results



i2b2-PHI	CVT(1b)	CVT(MIMIC)	Flair(1b)	Flair(MIMIC)	Bi-LSTM-CRF
Name	94.39	94.94	**95.68**	95.12	94.36
Age	94.94	**96.71**	95.32	95.95	95.19
Profession	77.06	**92.35**	81.47	81.17	77.35
Location	85.84	88.6,	89.17	**89.3**	87.17
Date	98.8	**98.98**	98.8	98.88	98.67
Contact	94.99	95.47	95.23	**95.7**	95.23
ID	90.94	91.12	**91.3**	90.5	90.23

Table 11: Recall (%) on i2b2-PHI categories on 2014 i2b2/UTHealth shared task Track 1. Best performance according to each category is highlighted.