# Applied Deep Learning

## Learning with Limited Labels for NLP

May 21st, 2019


YUN-NUNG (VIVIAN) CHEN

HTTP://ADL.MIULAB.TW

國立臺灣大學 National Taiwan University

# Outline

Limited Labeled Data

- How to incorporate the prior knowledge
- How to utilize the current observations

Unlabeled Data

- How to re-use the trained dialogue acts
- How to share knowledge across languages
- How to utilize parallel data

Conclusions

# Outline

**Limited Labeled Data**

- **How to incorporate the prior knowledge: Knowledge-Guided Model**
- How to utilize the current observations

Unlabeled Data

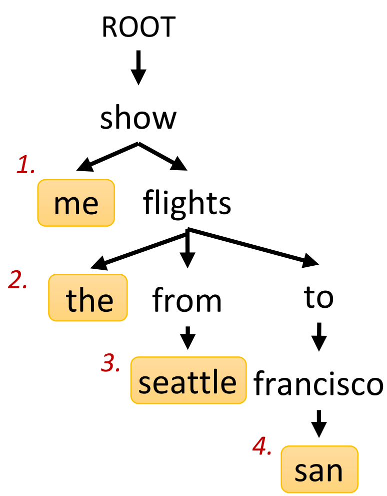
- How to re-use the trained dialogue acts
- How to share knowledge across languages
- How to utilize parallel data

Conclusions

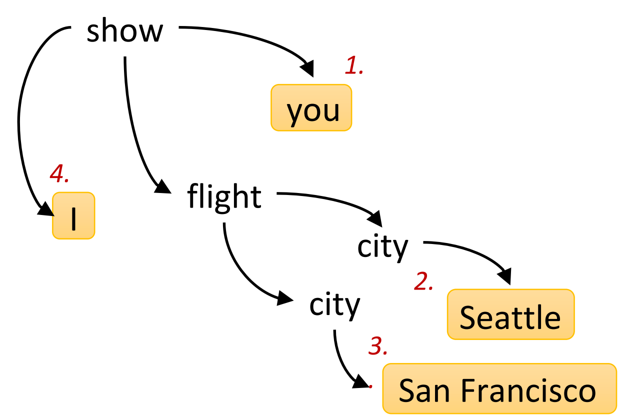
# Prior Structural Knowledge

**Sentence** $s$ show me the flights from seattle to san francisco

Syntax (Dependency Tree)




Semantics (AMR Graph)

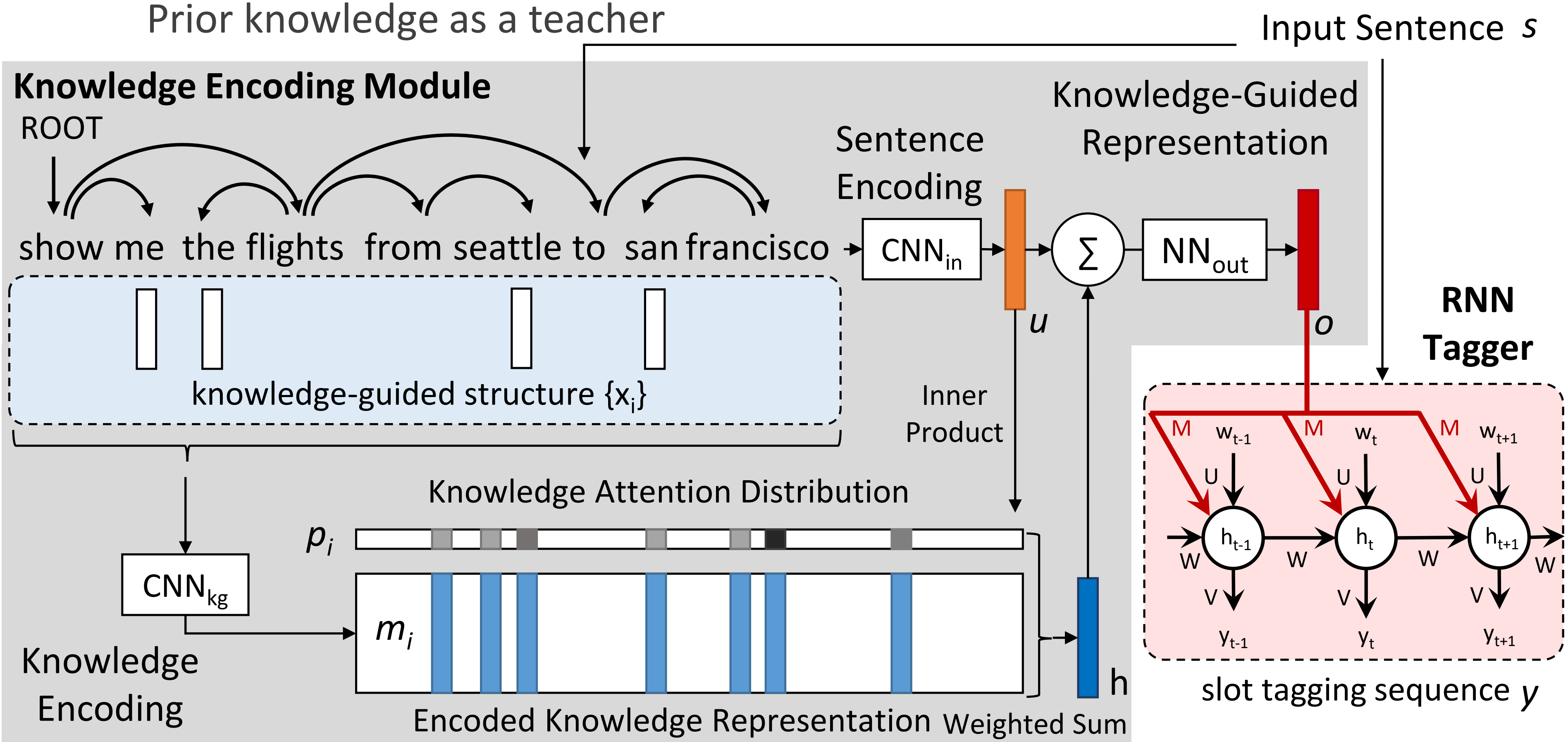



Prior knowledge about syntax or semantics may guide understanding

Y.-N. Chen, D. Hakkani-Tur, G. Tur, A. Celikyilmaz, J. Gao, and L. Deng, “Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks,” preprint arXiv: 1609.00777, 2016.

# K-SAN: Knowledge-Guided Structural Attention Networks

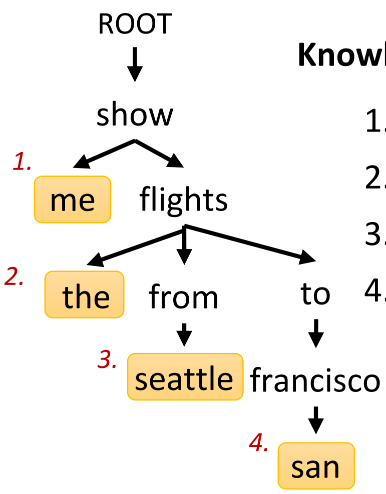
Prior knowledge as a teacher



# Sentence Structural Knowledge

**Sentence *s*** show me the flights from seattle to san francisco

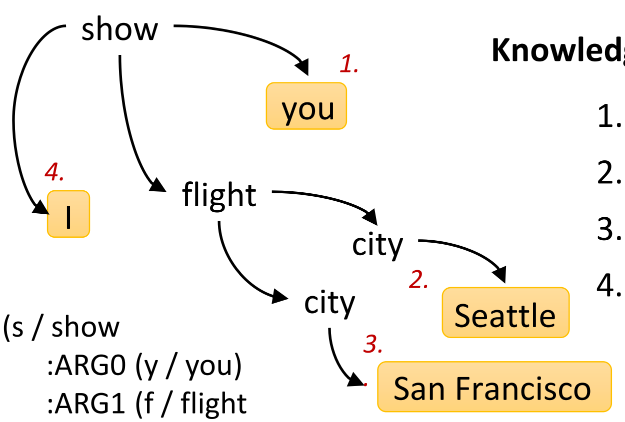
## Syntax (Dependency Tree)




**Knowledge-Guided Substructure $x_i$**

1. show me
2. show flights the
3. show flights from seattle
4. show flights to francisco san

## Semantics (AMR Graph)



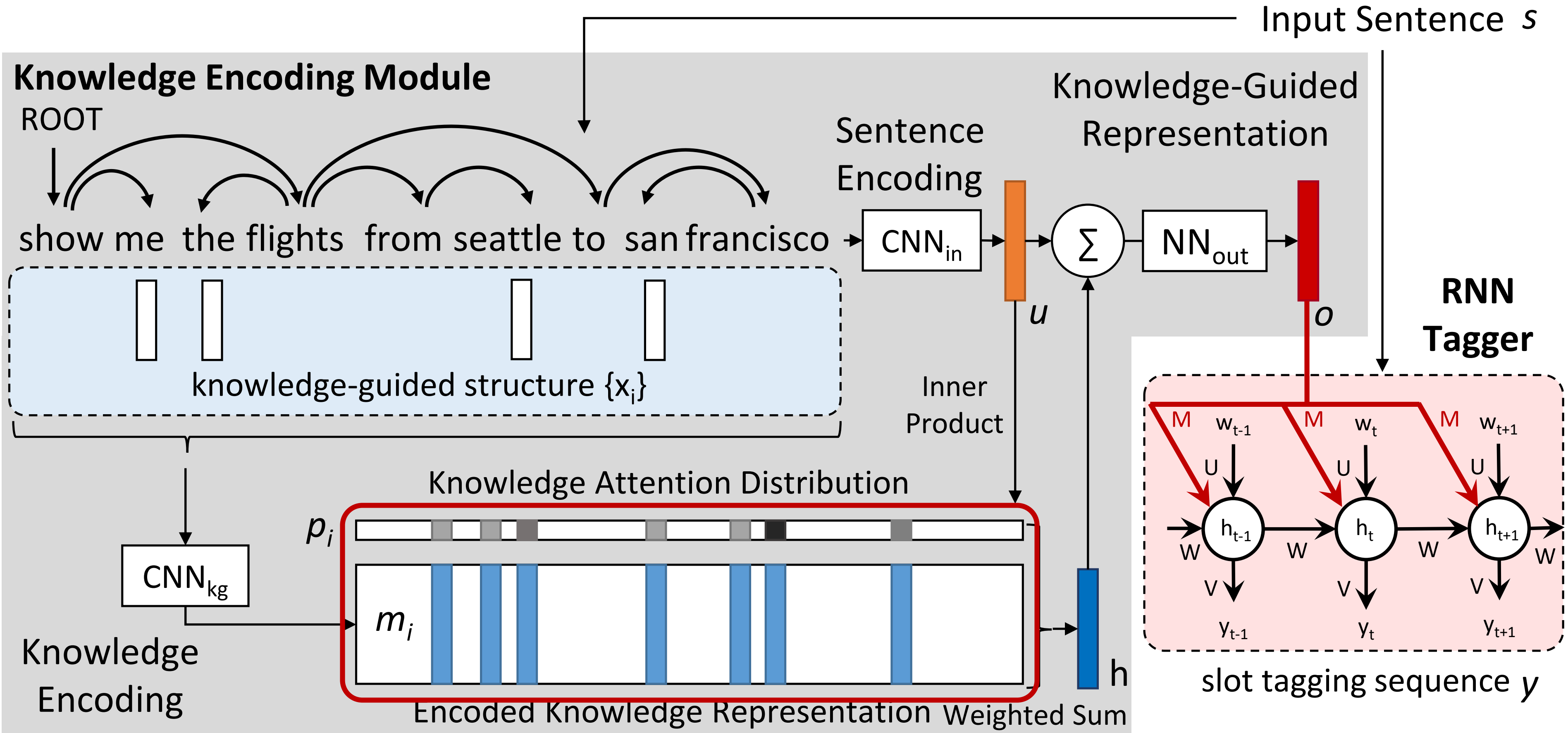

```
(s / show
    :ARG0 (y / you)
    :ARG1 (f / flight
        :source (c / city
            :name (d / name :op1 Seattle))
        :destination (c2 / city
            :name (s2 / name :op1 San :op2 Francisco)))
    :ARG2 (i / I)
    :mode imperative)
```

**Knowledge-Guided Substructure $x_i$**

1. show you
2. show flight seattle
3. show flight san francisco
4. show i

Y.-N. Chen, D. Hakkani-Tur, G. Tur, A. Celikyilmaz, J. Gao, and L. Deng, "Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks," preprint arXiv: 1609.00777, 2016.

# Knowledge-Guided Structures




The model will pay more attention to more important substructures that may be crucial for slot tagging.

Networks," preprint arXiv: 1609.00777, 2016.

# K-SAN Experiments

| ATIS Dataset (F1 slot filling) | Small (1/40) | Medium (1/10) | Large |
|---|---|---|---|
| Tagger (GRU) | 73.83 | 85.55 | 93.11 |
| Encoder-Tagger (GRU) | 72.79 | 88.26 | 94.75 |

# K-SAN Experiments

| ATIS Dataset (F1 slot filling) | Small (1/40) | Medium (1/10) | Large |
|---|---|---|---|
| Tagger (GRU) | 73.83 | 85.55 | 93.11 |
| Encoder-Tagger (GRU) | 72.79 | 88.26 | 94.75 |
| K-SAN (Stanford dep) | **74.60**+ | 87.99 | 94.86+ |
| K-SAN (Syntaxnet dep) | 74.35+ | **88.40**+ | **95.00**+ |

Syntax provides richer knowledge and more general guidance when less training data.

# K-SAN Experiments

| ATIS Dataset (F1 slot filling) | Small (1/40) | Medium (1/10) | Large |
|---|---|---|---|
| Tagger (GRU) | 73.83 | 85.55 | 93.11 |
| Encoder-Tagger (GRU) | 72.79 | 88.26 | 94.75 |
| K-SAN (Stanford dep) | **74.60**+ | 87.99 | 94.86+ |
| K-SAN (Syntaxnet dep) | 74.35+ | **88.40**+ | **95.00**+ |
| K-SAN (AMR) | 74.32+ | 88.14 | 94.85+ |
| K-SAN (JAMR) | 74.27+ | 88.27+ | 94.89+ |

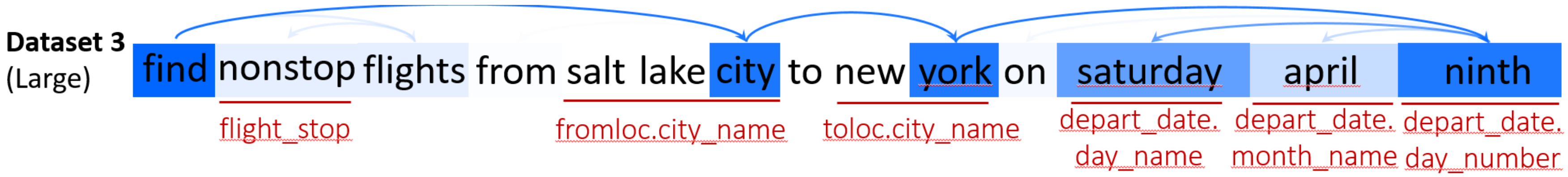
Syntax provides richer knowledge and more general guidance when less training data.

Semantics captures the most salient info so it achieves similar performance with much less substructures

Y.-N. Chen, D. Hakkani-Tur, G. Tur, A. Celikyilmaz, J. Gao, and L. Deng, "Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks," preprint arXiv: 1609.00777, 2016.

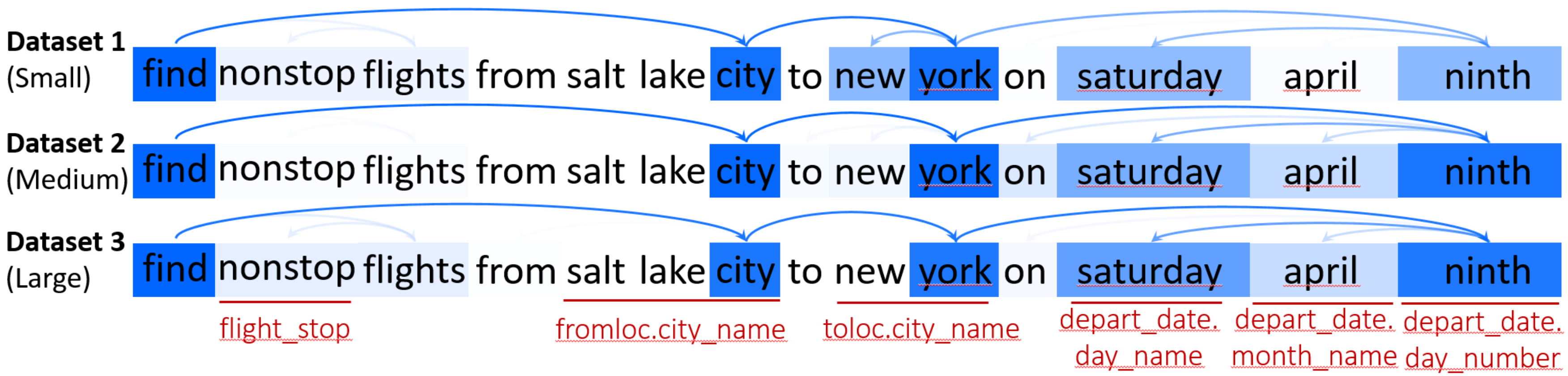
# Attention Analysis

Darker blocks and lines correspond to higher attention weights



# Attention Analysis

Darker blocks and lines correspond to higher attention weights




Using less training data with K-SAN allows the model pay the similar attention to the salient substructures that are important for tagging.

Y.-N. Chen, D. Hakkani-Tur, G. Tur, A. Celikyilmaz, J. Gao, and L. Deng, "Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks," preprint arXiv: 1609.00777, 2016.

# EHR Data

Predicting diagnosis codes for clinical reports

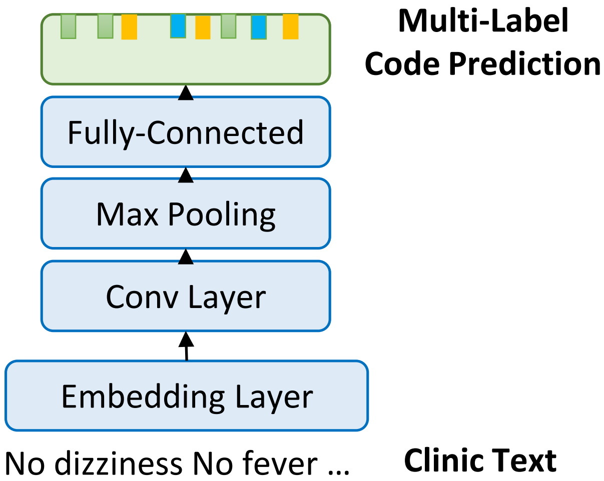
- Present illness text
  - “fever up to 39.4C intermittent in recent 3 days, cough/sputum(+), shortness of breath tonight”
- ICD-9 diagnosis codes
  - 486: Pneumonia, organism unspecified; 780.6: Fever

# CNN for Diagnosis Code Prediction

(Li et al., 2017)

Convolutional neural network (CNN) for multi-label code prediction

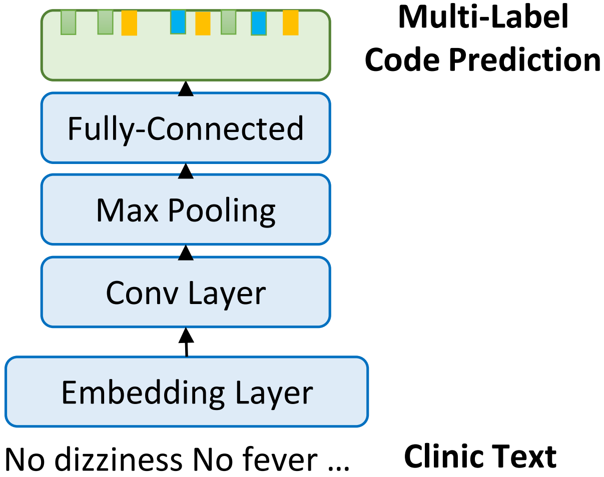
- Multiple convolutional filters for extracting different patterns




C. Li, et al., "Convolutional Neural Networks for Medical Diagnosis from Admission Notes," in arXiv, 2017.

# Hierarchy Category Knowledge

Idea: category knowledge provides additional cues to know code relatedness

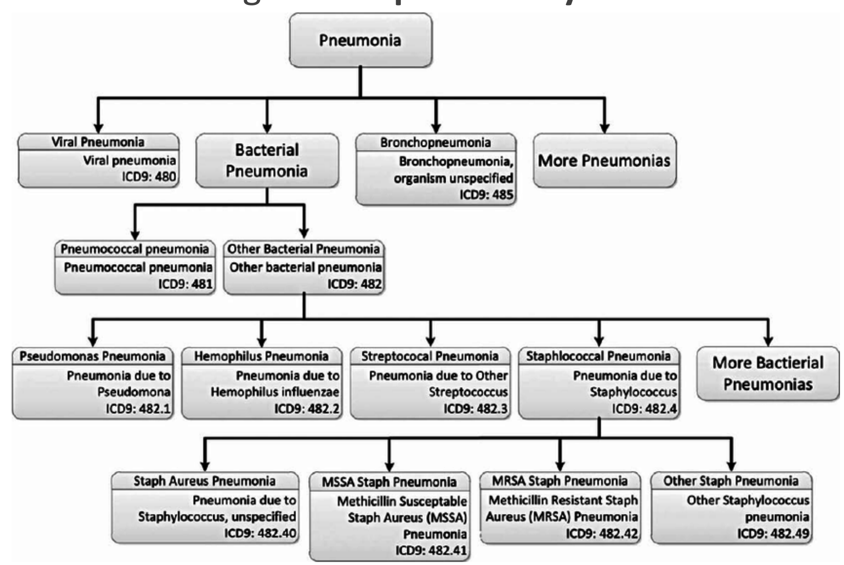



Low-level code

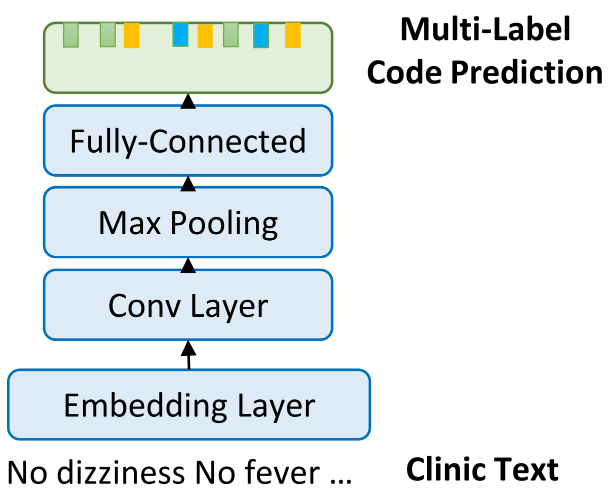
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category

- All belong to the “**personality disorders**”



# Hierarchy Category Knowledge
## (Cluster Penalty)




Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the "**personality disorders**"

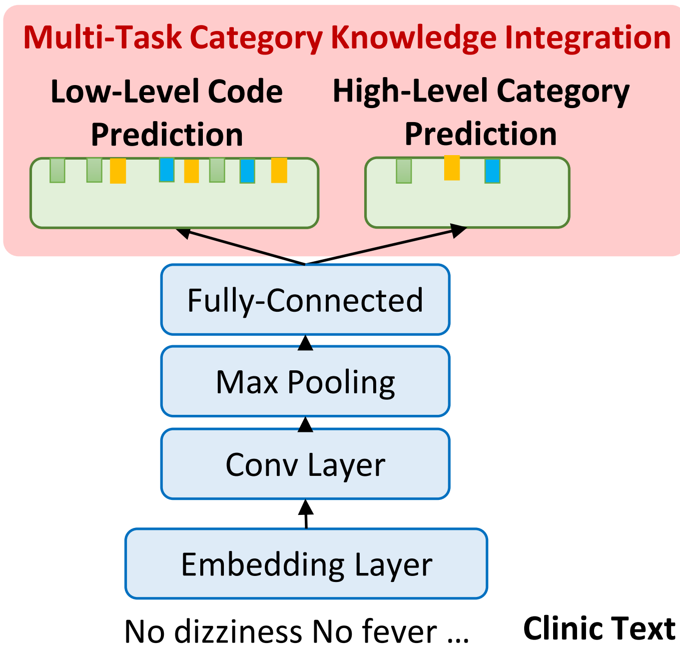
Category constrained loss

$$\Omega_{\text{between}} = \sum_{k=1}^{K} ||\bar{\theta}_k - \bar{\theta}||^2$$

$$\Omega_{\text{within}} = \sum_{k=1}^{K} \sum_{i \in \mathscr{J}(k)} ||\theta_i - \bar{\theta}_k||^2$$

A. Nie, et al., "DeepTag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain," in arXiv, 2018.

# Hierarchy Category Knowledge
## (Multi-Task)




Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the "**personality disorders**"

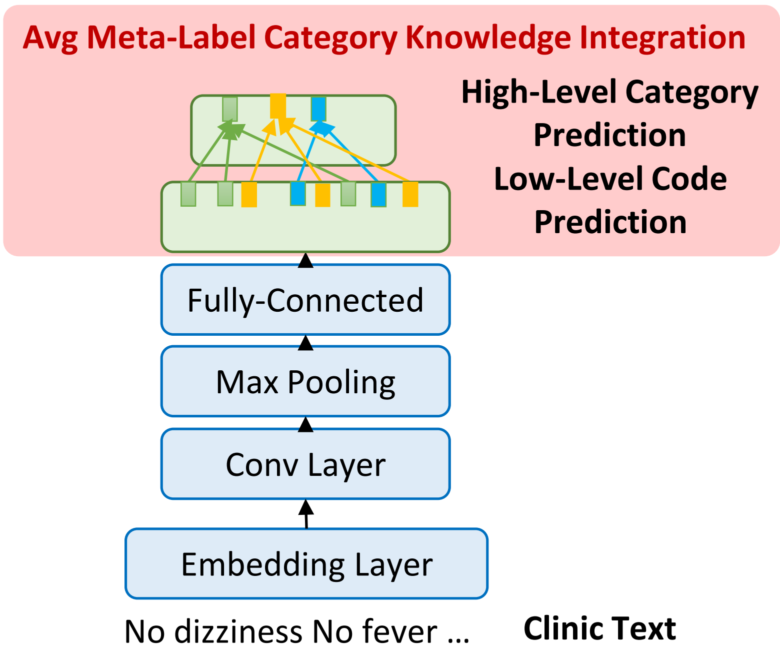
Low-level code infers the high-level category

$$y_{\text{high}} = 1 \text{ if } y_{\text{low}} = 1$$

Category integrated loss via multi-task

$$L = L_{\text{low}} + \gamma \cdot L_{\text{high}}$$

# Hierarchy Category Knowledge
## (Avg Meta-Label)




Low-level code

- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category

- All belong to the "**personality disorders**"

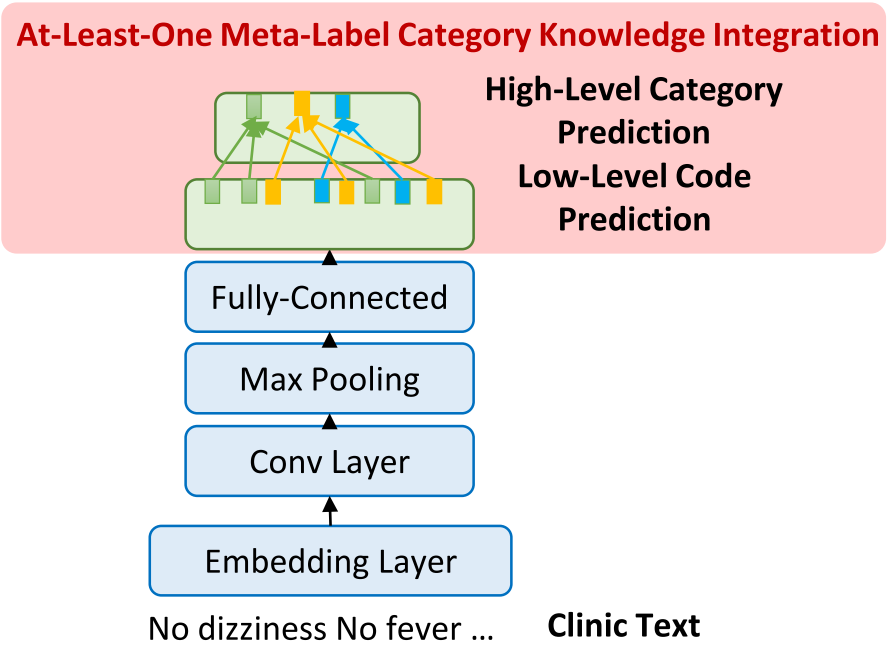
High-level prob can be approximated by the average of low-level code prob

$$y_{\text{high}} = \frac{1}{k} \sum y_{\text{low}}^{k}$$

Category integrated loss

$$L = L_{\text{low}} + \gamma \cdot L_{\text{high}}$$

# Hierarchy Category Knowledge (At-Least-One Meta-Label)




Low-level code

- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category

- All belong to the "**personality disorders**"

High-level prob can be approximated by the at-least-one of low-level code prob

$$y_{\text{high}} = 1 - \prod_k \left(1 - y_{\text{low}}^k\right)$$

Category integrated loss

$$L = L_{\text{low}} + \gamma \cdot L_{\text{high}}$$

# State-of-the-Art Performance

| MIMIC3-50 | | P@1 | P@3 | P@5 | MAP | Macro-F | Micro-F | Macro-AUC | Micro-AUC |
|---|---|---|---|---|---|---|---|---|---|
| CNN (Shi et al., 2017) | | 82.8 | 71.2 | 61.4 | 72.4 | 57.9 | 63.0 | 88.2 | 91.2 |
| + Cluster Penalty | | 83.5† | 71.9† | 62.4† | 73.1† | 58.3† | 63.7† | 88.5† | 91.3† |
| + Multi-Task | | 83.5† | 71.3† | 61.9† | 72.5† | 57.6 | 62.8 | 88.1 | 91.1 |
| + Hierarchical | *avg* | **84.5**† | **72.1**† | **62.4**† | **73.5**† | **58.6**† | **64.3**† | **88.9**† | **91.4**† |
| | *at-least-one* | 83.4† | 72.1† | 62.4† | 73.4† | 58.5† | 63.8† | 88.4† | 91.3† |
| **MIMIC3-Full** | | **P@1** | **P@3** | **P@8** | **P@15** | **Macro-F** | **Micro-F** | **Macro-AUC** | **Micro-AUC** |
| CNN (Shi et al., 2017) | | 80.5 | 73.6 | 59.6 | 45.4 | 3.8 | 42.9 | 81.8 | 97.1 |
| + Cluster Penalty | | 80.9† | 74.0† | 59.5 | 45.2 | 3.3 | 40.5 | 82.1† | 97.0 |
| + Multi-Task | | **82.8**† | **75.8**† | **61.5**† | **46.6**† | 3.6 | **43.9**† | **83.3**† | **97.3**† |
| + Hierarchical | *avg* | 79.0 | 73.1 | 59.2 | 45.2 | **4.3**† | 42.7 | 83.0† | 97.1 |
| | *at-least-one* | 82.1† | 74.3† | 59.7† | 44.9 | 2.6 | 42.0 | 80.3 | 96.7 |
| CAML (Mullenbach et al., 2018) | | 89.6 | 83.4 | 69.5 | 54.6 | 6.1 | 51.7 | 88.4 | 98.4 |
| + Cluster Penalty | | 88.4 | 82.4 | 68.8 | 54.0 | 5.4 | 51.2 | 87.5 | 98.3 |
| + Multi-Task | | **89.7**† | 83.4 | 69.7† | 54.8 | 6.9† | 52.3† | 88.8† | 98.5† |
| + Hierarchical | *avg* | 89.6 | **83.5**† | **70.9**† | **56.1**† | **8.2**† | **53.9**† | **89.5**† | **98.6**† |
| | *at-least-one* | 89.4 | 83.3 | 69.5 | 54.8† | 6.2† | 51.7 | 88.3 | 98.4 |

# Outline

**Limited Labeled Data**

- How to incorporate the prior knowledge: Knowledge-Guided Model
- **How to utilize the current observations: Semi-Supervised Multi-Task SLU**

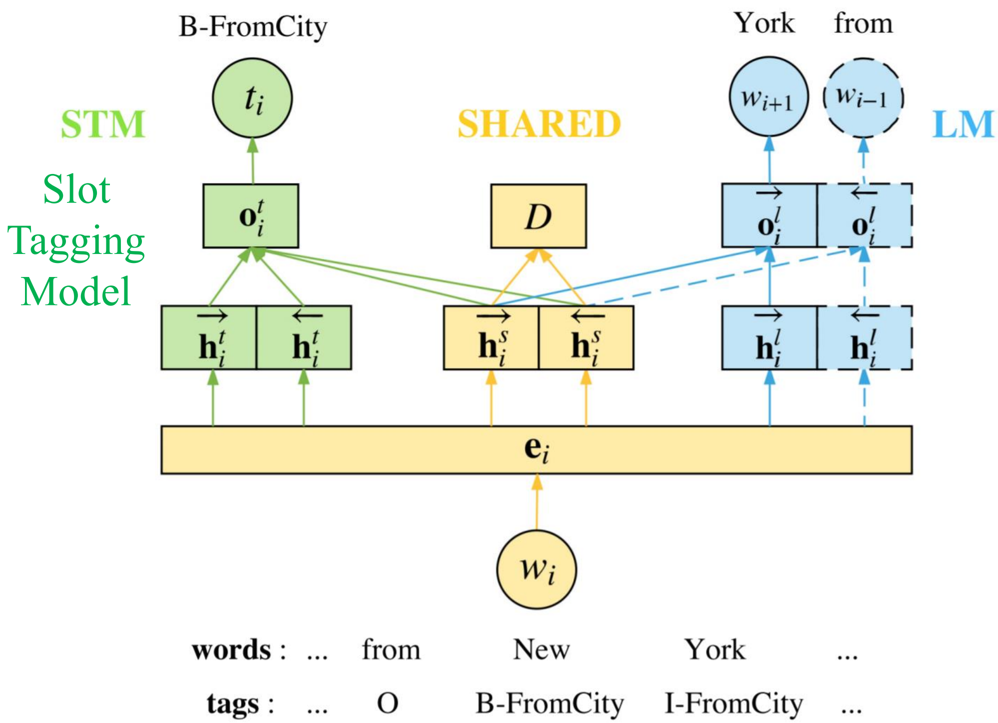
Unlabeled Data

- How to re-use the trained dialogue acts
- How to share knowledge across languages
- How to utilize parallel data

Conclusions

# Semi-Supervised Multi-Task SLU (Lan et al., 2018)

Idea: language understanding objective can enhance other tasks




**Algorithm 1:** Adversarial Multi-task Learning for SLU

**Input** : Labeled training data $\{(\mathbf{w}^l, \mathbf{t}^l)\}$
Unlabeled data $\{\mathbf{w}^u\}$

**Output:** Adversarially enhanced slot tagging model

Initialize parameters $\{\theta^s, \theta^t, \theta^l, \theta^d\}$ randomly.

**repeat**

```
/* Sample from {(w^l, t^l)} */
```

Train the STM and shared model by Eq.(8).

Train the task discriminator and the shared model by Eq.(6) or Eq.(7) as slot tagging task ($y = 1$).

```
/* Sample from {w^l} and {w^u} */
```

Train the LM and shared models by Eq.(9) (and Eq.(10) for BLM).

Train the task discriminator and the shared model by Eq.(6) or Eq.(7) as LM task ($y = 0$).

**until** *convergence*;

BLM exploits the *unsupervised knowledge*, the *shared-private framework* and *adversarial training* make the slot tagging model more generalized

O. Lan, S. Zhu, and K. Yu, "Semi-supervised Training using Adversarial Multi-task Learning for Spoken Language Understanding," in *Proceedings of ICASSP*, 2018.

# Semi-Supervised Multi-Task SLU (Lan et al., 2018)

STM – BLSTM for slot tagging

MTL – multi-task learning for STM and LM, where they share the embedding layer

PSEUDO – train an STM with labeled data, generate labels for unlabeled data, and retrain STM

| Method | 5k | 10k | 15k | all |
|---|---|---|---|---|
| STM | 67.25 | 71.04 | 73.94 | 76.60 |
| $\text{MTL}_e$ | 69.57 | 73.04 | 75.00 | 77.24 |
| PSEUDO | 69.82 | 72.55 | 74.80 | - |
| BSPM | 68.46 | 72.52 | **75.05** | **77.52** |
| $\text{BSPM+D}^{(w)}$ | **71.55** | **73.67** | 74.61 | 77.42 |
| $\text{BSPM+D}^{(s)}$ | 70.99 | 73.58 | 74.22 | 77.24 |

The model is more efficient when the labeled data is limited and the data for LM is more sufficient.

O. Lan, S. Zhu, and K. Yu, "Semi-supervised Training using Adversarial Multi-task Learning for Spoken Language Understanding," in *Proceedings of ICASSP*, 2018.

# Outline

Limited Labeled Data

- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

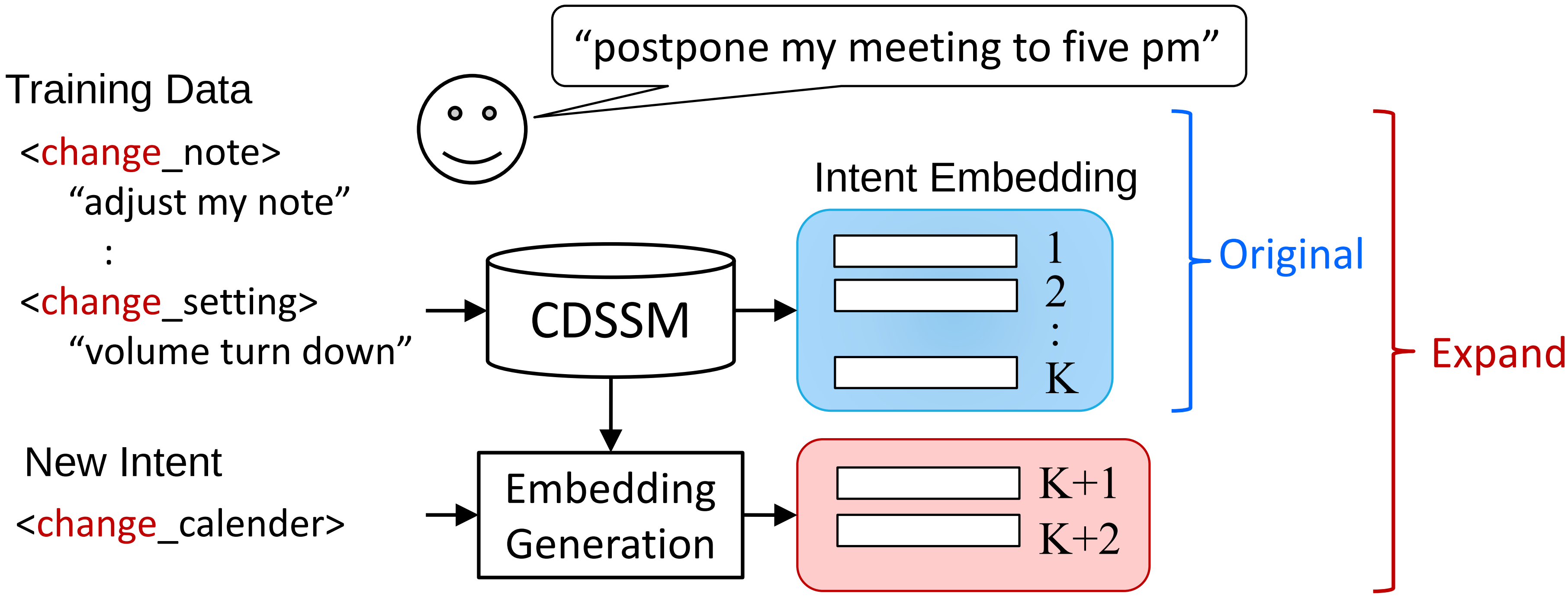
**Unlabeled Data**

- **How to re-use the trained dialogue acts: Zero-Shot Intent Expansion**
- How to share knowledge across languages
- How to utilize parallel data

Conclusions

# Zero-Shot Intent Expansion (Chen et al., 2016)

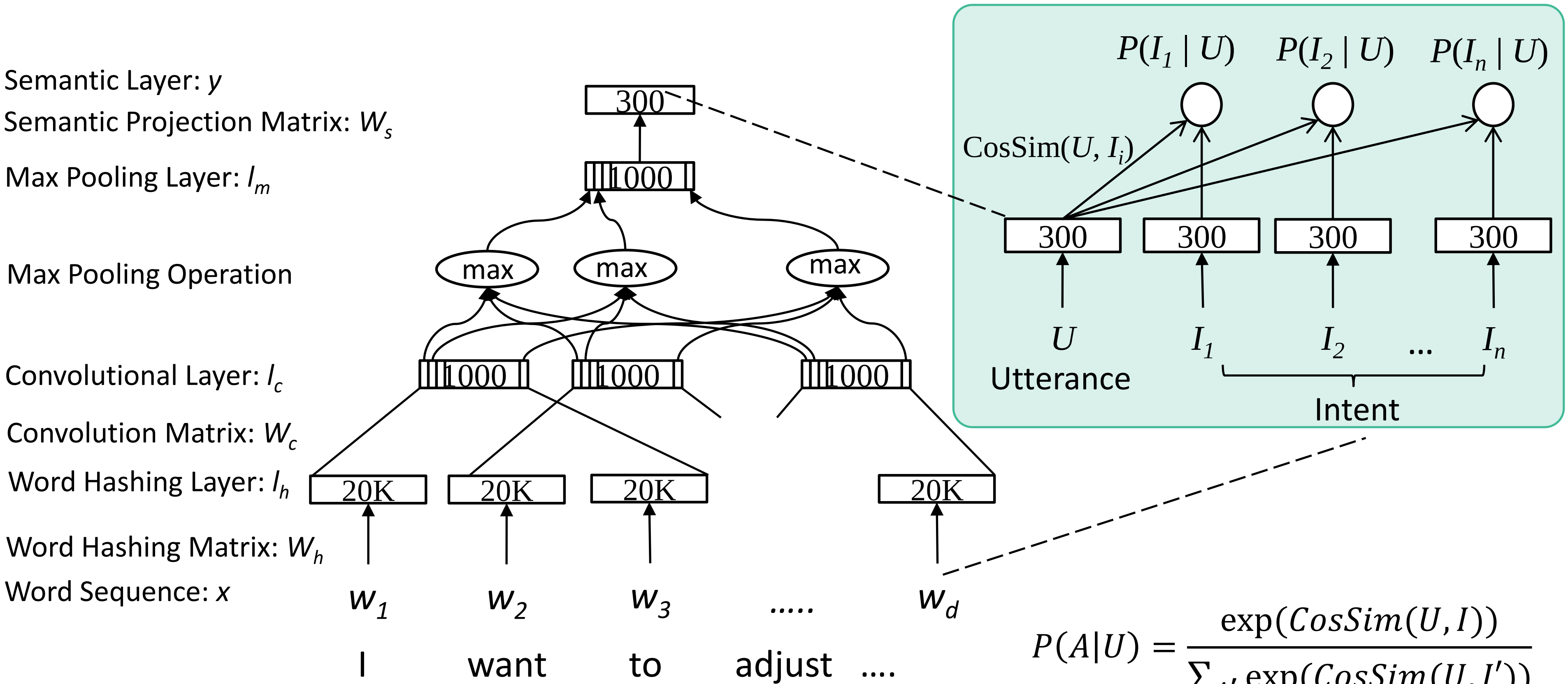
Goal: resolve domain constraint and enable flexible intent expansion for unlabeled domains




Same dialogue acts can be shared across domains


Y.-N. Chen, D. Hakkani-Tur, and X. He, "Zero-Shot Learning of Intent Embeddings for Expansion by Convolutional Deep Structured Semantic Models," in *Proceedings of ICASSP*, 2016.

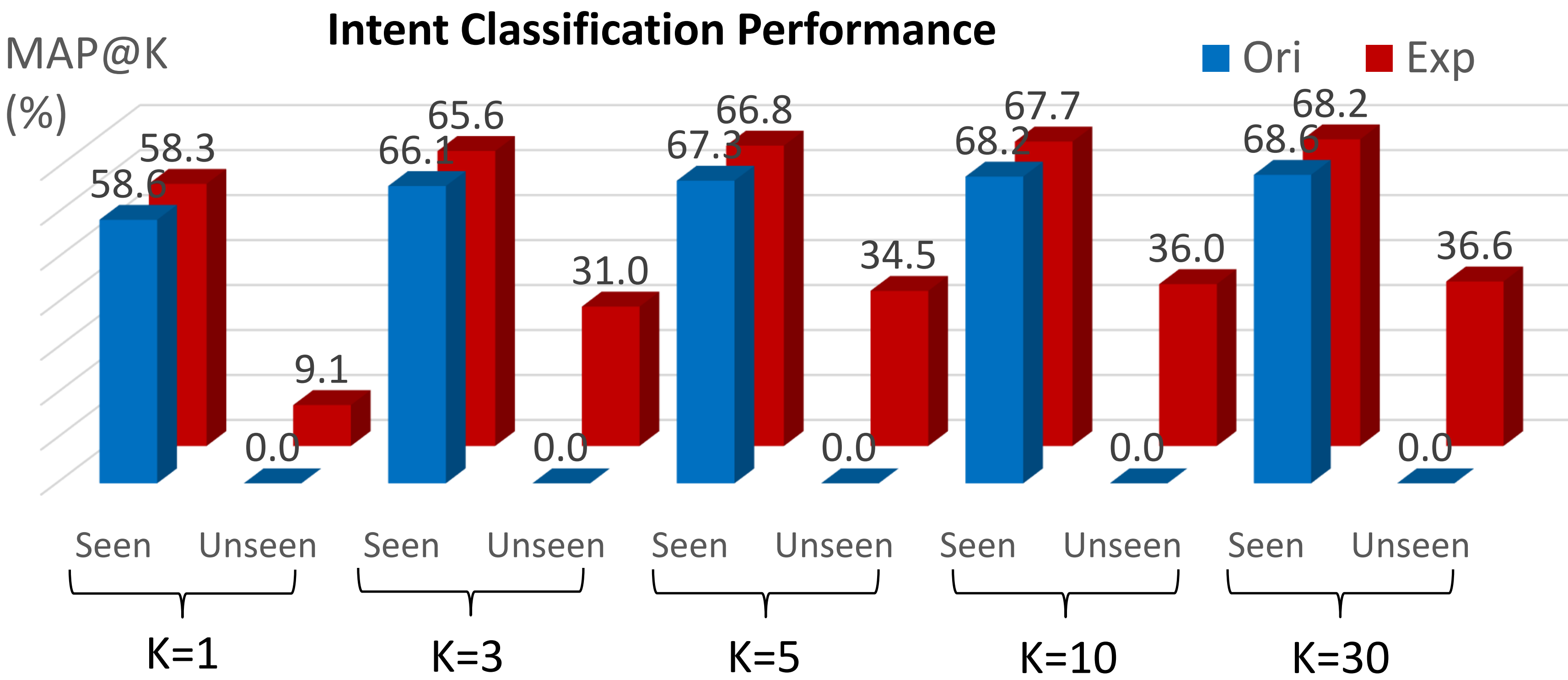
# CDSSM: Convolutional Deep Structured Semantic Models




$$P(A|U) = \frac{\exp(CosSim(U,I))}{\sum_{A'} \exp(CosSim(U,I'))}$$

CDSSM maps language usage for the same dialogue acts together

# Zero-Shot Intent Expansion (Chen et al., 2016)




The expanded models consider new intents without training samples, and produces better understanding for unseen domains with comparable results for seen domains.

# Outline

Limited Labeled Data

- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

**Unlabeled Data**

- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- **How to share knowledge across languages: Zero-Shot Crosslingual SLU**
- How to utilize parallel data

Conclusions

# Zero-Shot Crosslingual SLU (Upadhyay et al., 2018)

Source language: English (full annotations)

Target language: Hindi (limited annotations)

RT: round trip, FC: from city, TC: to city, DDN: departure day name

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Utt: | find | a | one | way | flight | from | boston | to | atlanta | on | wednesday |
| Slots: | O | O | B-RT | I-RT | O | O | B-FC | O | B-TC | O | B-DDN |

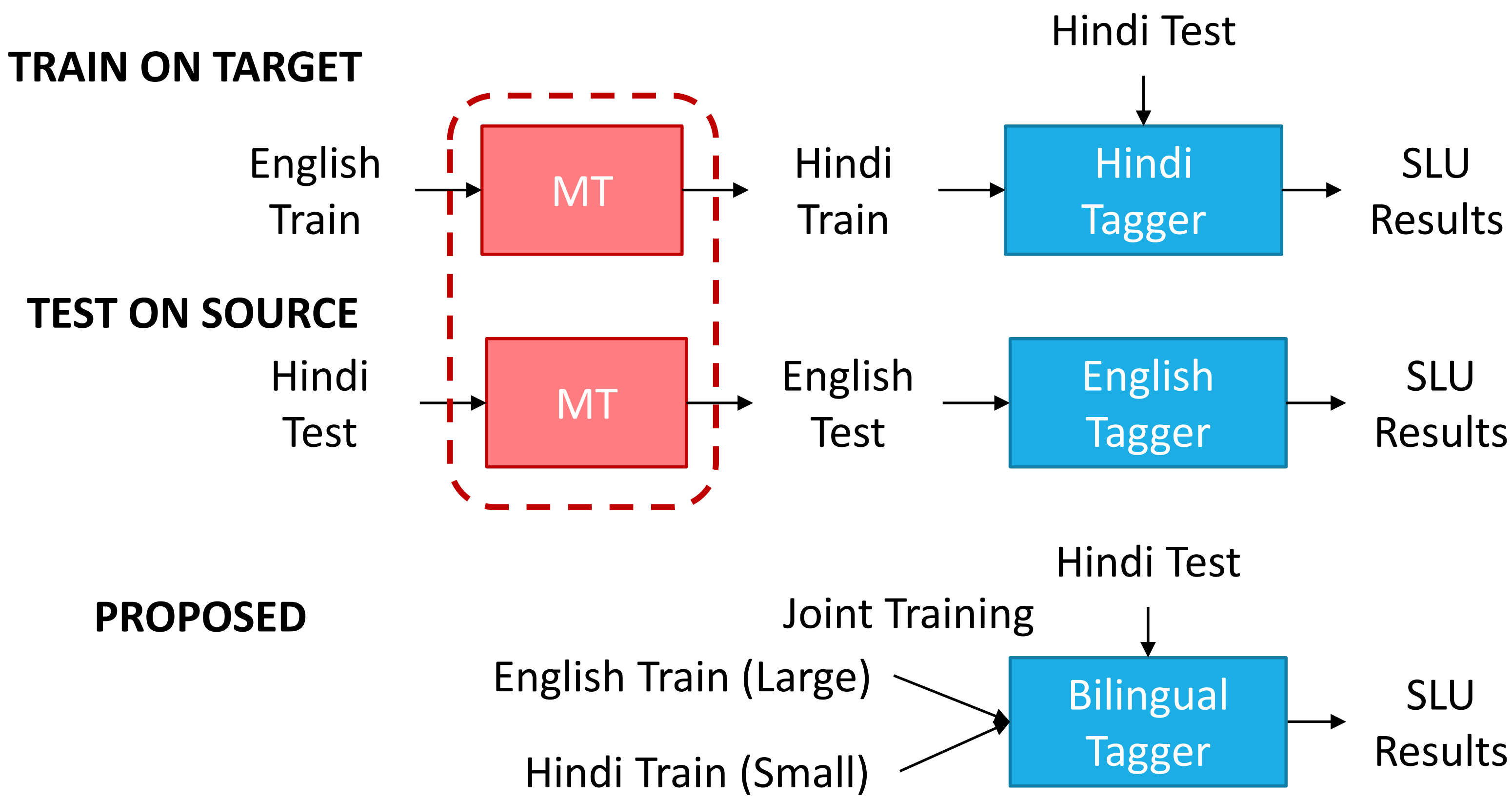
(a) English Utterance

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Utt: | बुधवार | को | बोसटन | से | अटलांटा | तक | जाने | वाली | एकतरफ़ा | उड़ाने | खोजें |
| Slots: | B-DDN | O | B-FC | O | B-TC | O | O | O | B-RT | O | O |

(b) Hindi Utterance

S. Upadhyay, M. Faruqui, G. Tur, D. Hakkani-Tur, and L. Heck, "(Almost) Zero-Shot Cross-Lingual Spoken Language Understanding," in *Proceedings of ICASSP*, 2018.

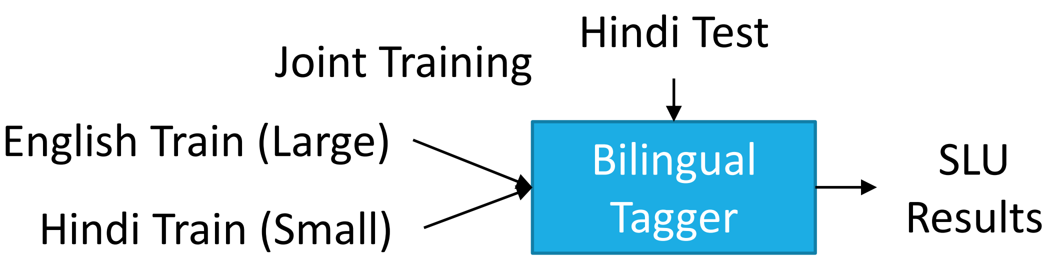
# Zero-Shot Crosslingual SLU (Upadhyay et al., 2018)

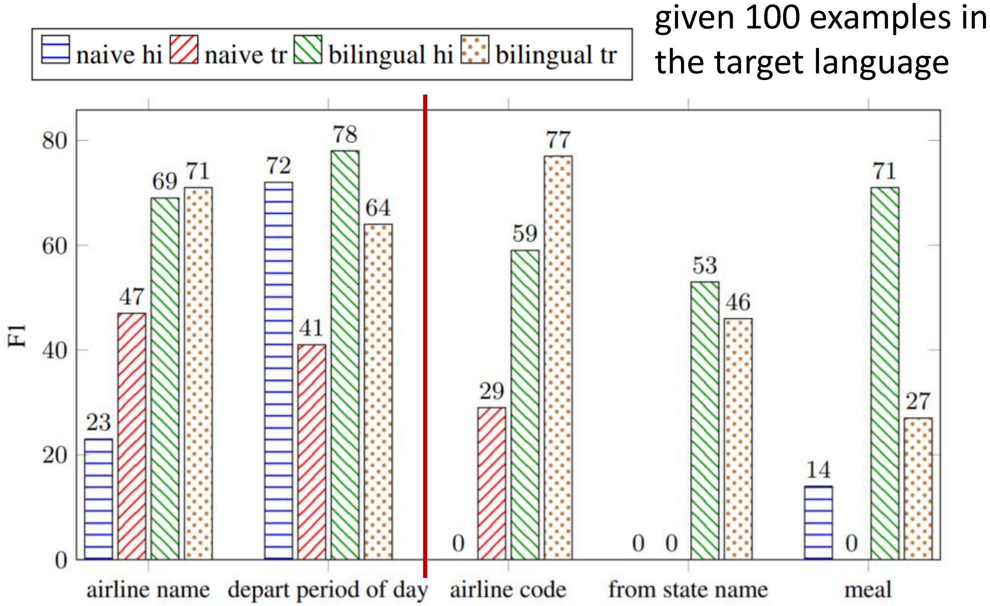



MT system is not required and both languages can be processed by a single model

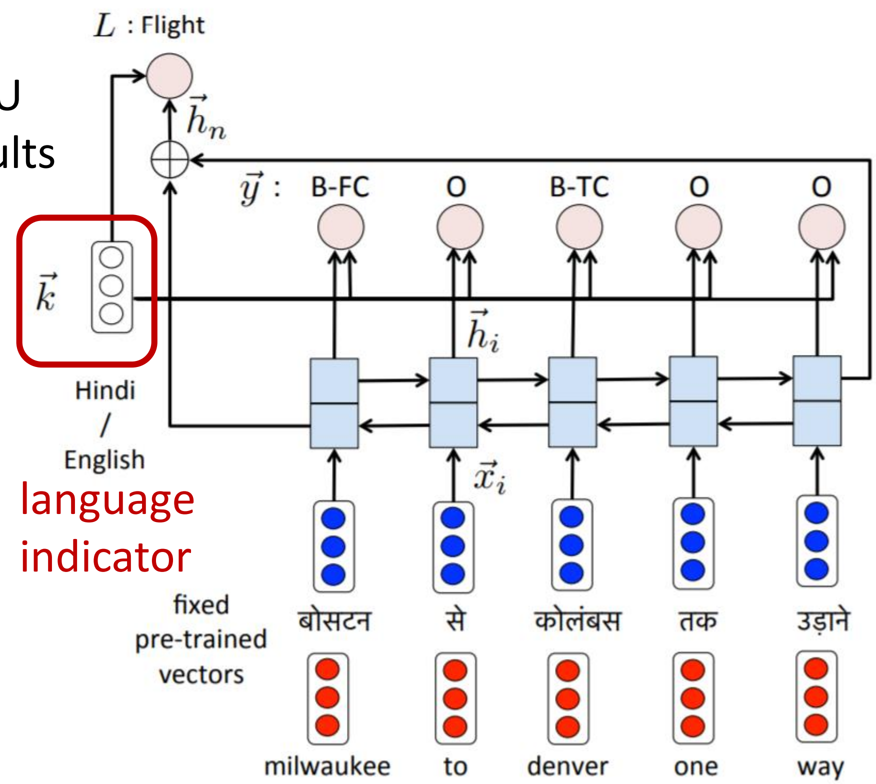
S. Upadhyay, M. Faruqui, G. Tur, D. Hakkani-Tur, and L. Heck, "(Almost) Zero-Shot Cross-Lingual Spoken Language Understanding," in *Proceedings of ICASSP*, 2018.

# Joint Model for Crosslingual SLU



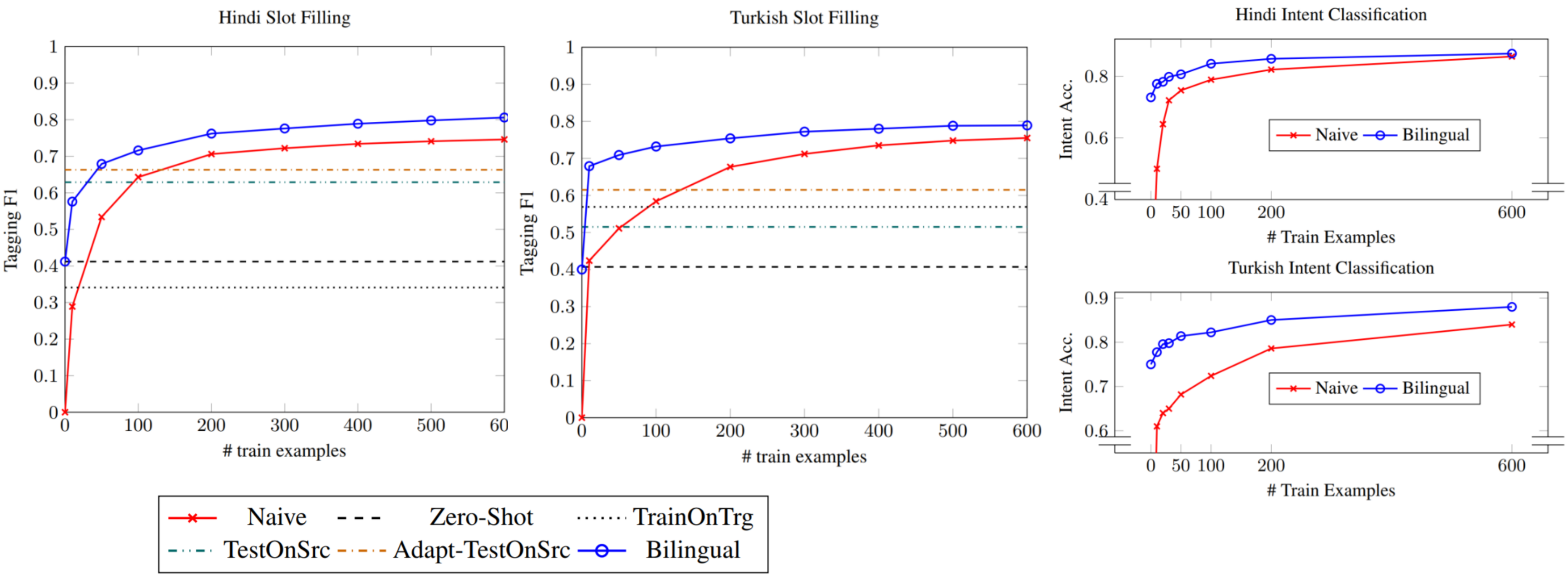




given 100 examples in the target language




For rare slots (like meal, airline code), there is a huge difference between the bilingual model and the naive model when the target training data is limited

# Bilingual Model SLU Experiments




The bilingual model outperforms others and does not suffer from latency introduced by MT


S. Upadhyay, M. Faruqui, G. Tur, D. Hakkani-Tur, and L. Heck, "(Almost) Zero-Shot Cross-Lingual Spoken Language Understanding," in *Proceedings of ICASSP*, 2018.

# Outline

Limited Labeled Data

- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

**Unlabeled Data**

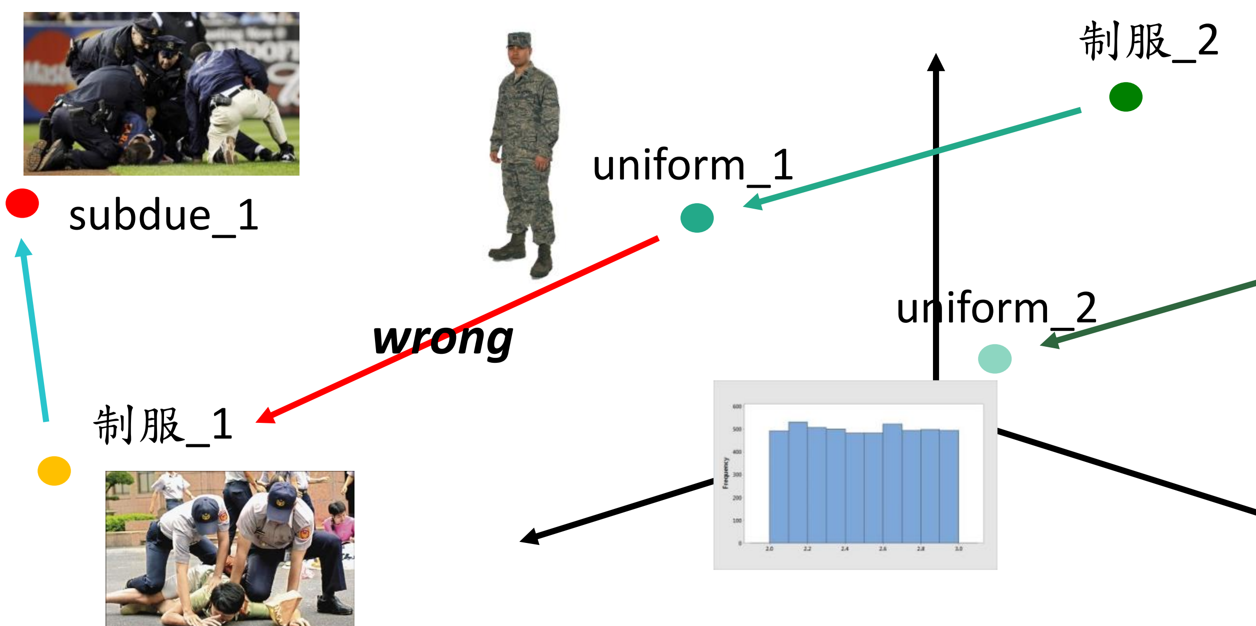
- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages: Zero-Shot Crosslingual SLU
- **How to utilize parallel data: Crosslingual Sense Embeddings**

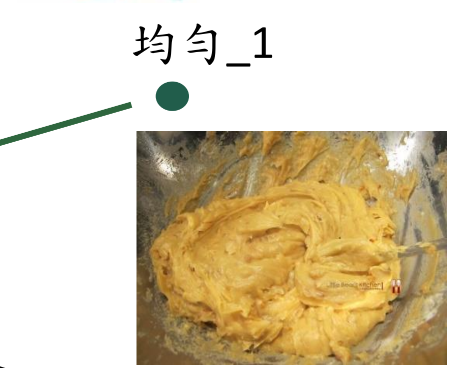
Conclusions

# Crosslingual Embeddings

Tokens in source language shall be mapped to tokens in target language

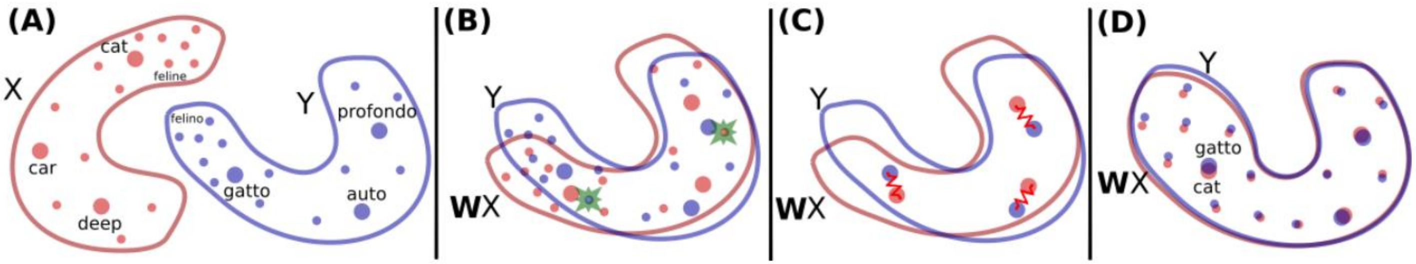
- This assumption only holds in sense level token
- Sets of crosslingual sense embeddings are therefore important
- uniform/制服 are all polysemous words

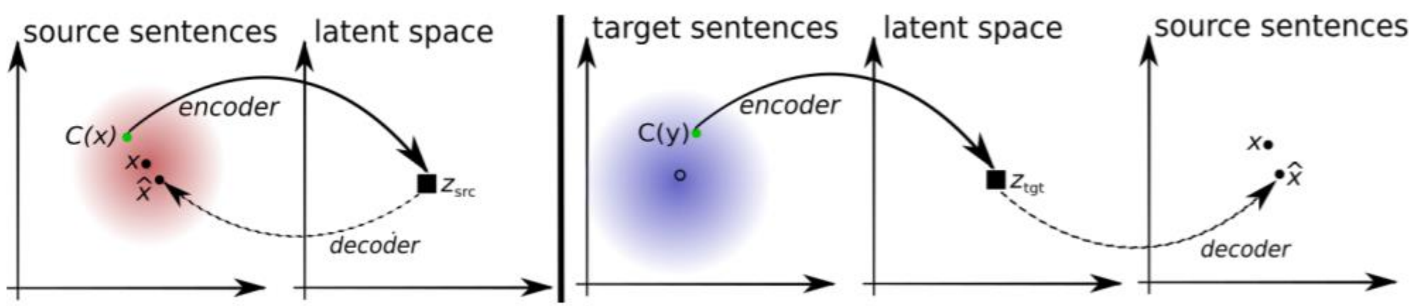



# Embeddings in a Unified Space (Conneau et al., 2017; Lample et al., 2017)

May largely benefit tasks such as unsupervised machine translation

- 


- 


A. Conneau, G. Lample, L. Denoyer, MA. Ranzato, H. Jégou, "Word Translation Without Parallel Data," *preprint arXiv: 1710:04087*, 2017.
G. Lample, A. Conneau, L. Denoyer, MA. Ranzato, "Unsupervised Machine Translation With Monolingual Data Only," *preprint arXiv:1711.00043*, 2017.

# Modular Framework

Our method can be separated into two steps (Lee & Chen, 2017):

1. Select the most probable (argmax) sense given the context
2. Use skip-gram to train the representation of the selected senses

➢ Reinforcement learning is used to connected the two modules


Lee and Chen, "MUSE: Modularizing Unsupervised Sense Embeddings," in *EMNLP*, pages 327-337, 2017.

# Sense Selection Module

Input:

- Chinese text context $\mathrm{C_t} = [C_{t-m}, \dots, C_t = w_i, \dots, C_{t+m}]$
- English text context $\mathrm{C_t}' = [C'_{t-m}, \dots, C'_t = w'_i, \dots, C'_{t+m}]$

Output: the fitness for each sense $z_{i1}, \dots, z_{i3}$

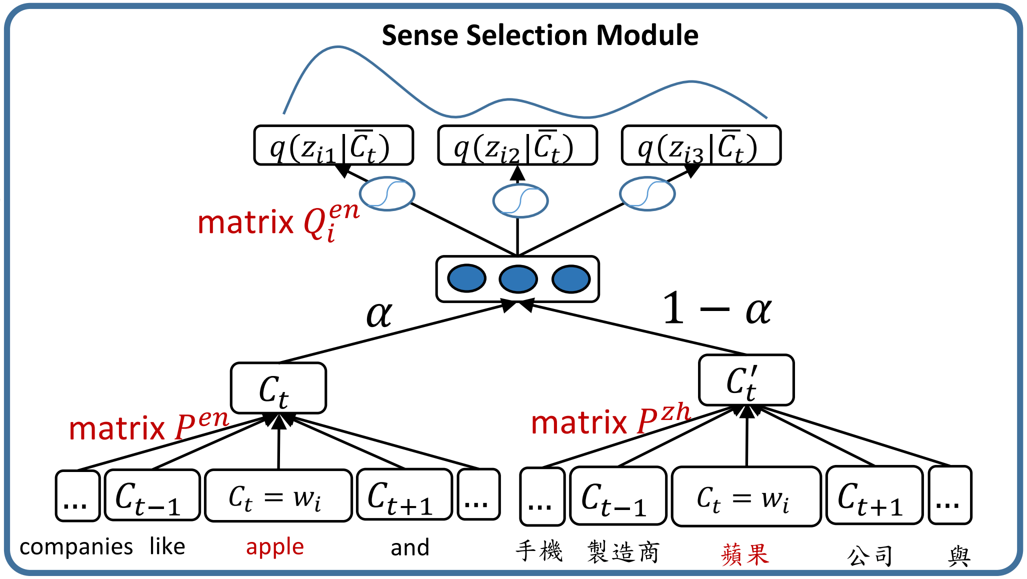
Model architecture: Continuous Bag-of-Words (CBOW) for efficiency

Sense selection

$$\bar{C} = \alpha \cdot \frac{1}{|c_i|} \sum_{w_j \in c_i} P_j^{en} + (1-\alpha) \cdot \frac{1}{M} \sum_{w'_j \in c'_i} P_j^{zh}$$

$$p(z_{ik} \mid c_i, c'_i) = \sigma((Q_{ik}^{en})^T \bar{C})$$

$$z_{ik}^* = \arg\max_{z_{ik}} p(z_{ik} \mid c_i, c'_i)$$

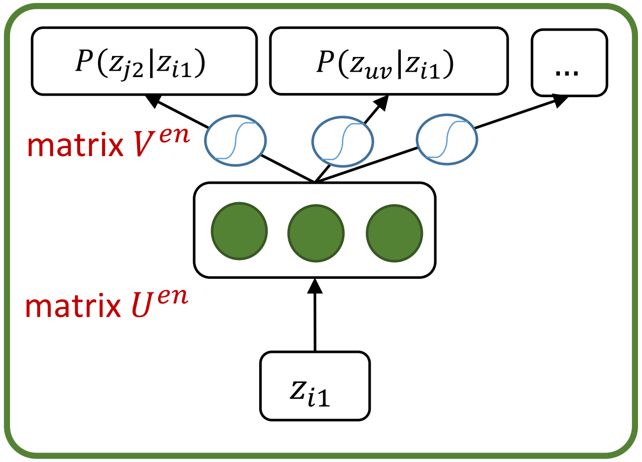


# Sense Representation Module

Input: sense collocation $s_i, s_j, s_l'$

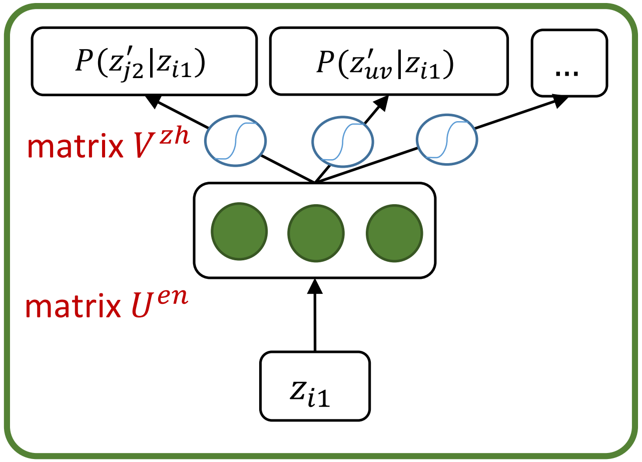
Output: collocation likelihood estimation

Model architecture: skip-gram architecture

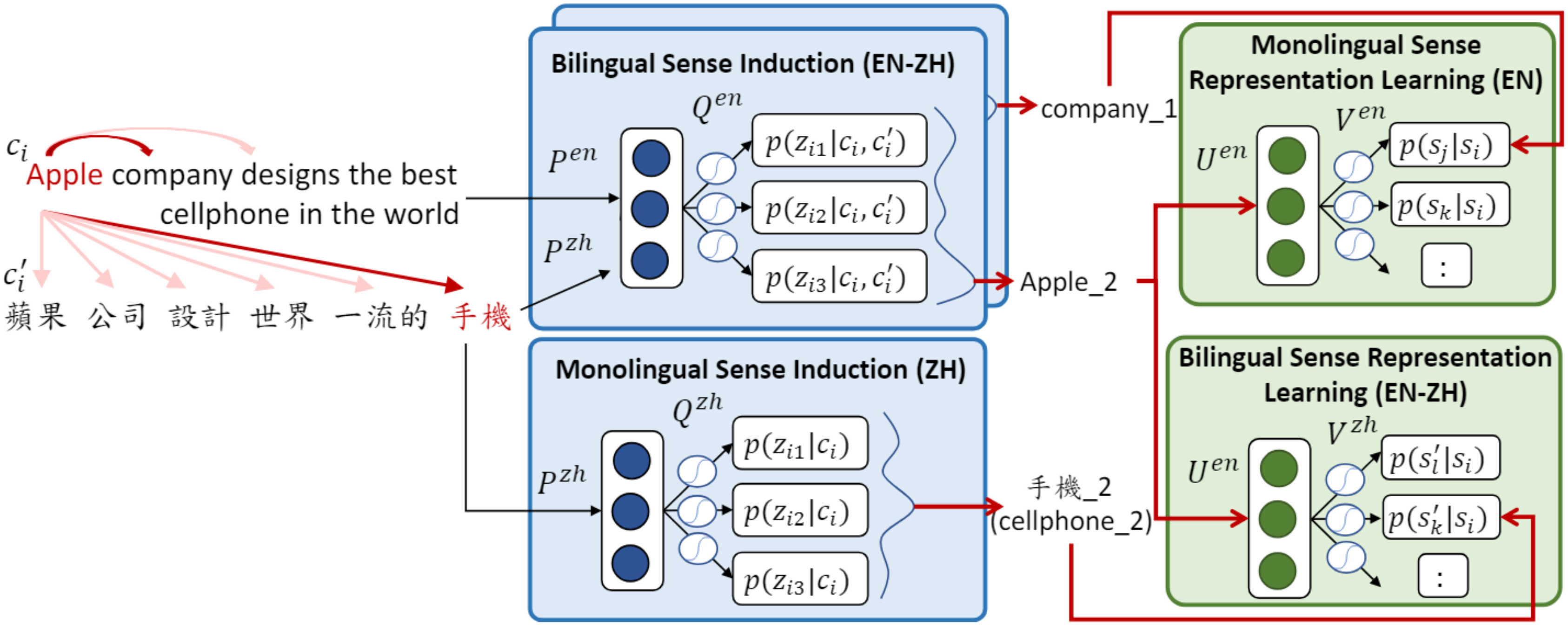



Sense selection (optimized by negative sampling)

$$p(s_l' \mid s_i) = \frac{\exp((U_{s_i}^{en})^T V_{s_l'}^{zh})}{\sum_{s_k'} \exp((U_{s_i}^{en})^T V_{s_k'}^{zh})}$$



# Crosslingual Model Architecture




Enabling bilingual sense embedding learning with parallel data

# Qualitative Analysis

| Target | kNN Senses (EN) | kNN Senses (ZH) |
|---|---|---|
| apple_0 | fruit, cake, sweet | 蘋果, 春天, 蛋糕, iphone, 雞蛋, 巧克力, 葡萄<br>(apple, spring, cake, iphone, egg, chocolate, purples) |
| apple_1 | iphone, cake, google, stores | 蘋果, iphone, 微軟, 競爭對手, 春天, 谷歌<br>(apple, iphone, microsoft, competitor, spring, google) |
| uniform_0 | dressed, worn, tape, wearing, cloth | 均勻,光滑,衣服,鞋子,穿著,服裝<br>(even, smooth, clothes, shoes, wearing, clothing) |
| uniform_1 | particle, computed, varying, gradient | 態,粉末,縱向,等離子體,剪切,剛度<br>(phase, powder, longitudinal, plasma, cut, stiffness) |

The words with similar senses from both languages have similar embeddings in a unified space

# New Dataset – BCWS
## (Bilingual Contextual Word Similarity)

| English sentence | Chinese sentence | Score |
|---|---|---|
| Judges must give both sides an equal opportunity to <**state**> their cases. | 我非常喜歡這個故事，它<**告訴**>我們一些重要的啓示。 (I like this story a lot, which <**tells**> us some important inspiration.) | 7.00 |
| It was of negligible <**importance**> prior to 1990, with antiquated weapons and few members. | 黃斑部病變的預防及早期治療是相當<**重要**>的。 (The prevention and early treatment of macular lesions is very <**important**>.) | 6.94 |
| Due to the San Andreas Fault bisecting the hill, one side has <**cold**> water, the other has hot. | 水果攤老闆似乎很意外眞有人買這<**冷**>貨，露出「你眞內行」的眼神與我聊了幾句。 (The owner of the fruit stall seemed surprised that someone bought this <**unpopular**> product, talking me few words about "you are such a pro".) | 3.70 |

A newly collected dataset for evaluating bilingual sense embeddings

# Contextual Word Similarity Experiment

| Model | $\alpha$ | EN-ZH | | EN-DE |
|---|---|---|---|---|
| | | Bilingual/BCWS | Mono(EN)/SCWS | Mono(EN)/SCWS |
| *1) Monolingual Sense Embeddings* | | | | |
| Lee and Chen (2017) | | | **66.8** / 65.5 | **63.8** / 63.4 |
| *2) Crosslingual Word Embeddings* | | | | |
| Luong et al. (2015) | | 49.2 | 61.1 | 62.1 |
| Conneau et al. (2017) | | 52.5 | 65.5 | **64.0** |
| *3) Crosslingual Sense Embeddings* | | | | |
| Upadhyay et al. (2017) | | - | 45.0[2] | - |
| Proposed | 0.1 | 55.8 / 55.8 | 65.6 / 65.6 | **63.8** / 63.9 |
| | 0.3 | 55.7 / 55.7 | 64.9 / 65.1 | **63.8 / 64.0** |
| | 0.5 | 56.3 / 56.3 | 65.8 / 66.0 | 63.6 / 63.9 |
| | 0.7 | **56.7 / 56.7** | 65.6 / 65.8 | 63.1 / 63.2 |
| | 0.9 | 56.0 / 56.0 | 66.0 / **66.2** | 62.9 / 63.1 |

The crosslingual sense embeddings learned in an unsupervised way produce better results on BCWS (bilingual) and comparable performance on SCWS (monolingual)

# Outline

## Limited Labeled Data

- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

## Unlabeled Data

- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages: Zero-Shot Crosslingual SLU
- How to utilize parallel data: Crosslingual Sense Embeddings

## **Conclusions**

# Concluding Remarks

*Prior knowledge* can benefit understanding when less training data

*Language modeling objective* can be incorporated to benefit other tasks

*Dialogue acts* can be shared across different domains

*Crosslingual word embeddings* and *joint model* help extend models to different languages

Sense-level representations can be learned via *contexts*

The *parallel data* for MT can bridge the embeddings from different languages