苏州大学

Soochow University

Conversion of Multiple Resources for POS Tagging

Enting Gao[1], Jiayuan Chao[2], Zhenghua Li[2]

1、Suzhou University of Science and Technology， suzhou, China

2、Soochow University, suzhou, China

Outline

- Background
- Related Work
- Motivation
- Our Method
- Experiments
- Conclusion

Background

Definition: Part-of-speech Tagging

- Mark up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context.

Example

Input	外交部$_1$	发言人$_2$	答$_3$	记者$_4$	问$_5$
	Foreign Ministry	spokesman	answers	reporters'	questions
Output	NR	NN	VV	NN	NN

Related work

Recent advance

- Tseng et al. (2005) propose a variety of **morphological features** for POS tagging to help improve unknown-word tagging performance
- Huang et al. (2009) utilize a **discriminative reranking model** to help achieve improvement on Mandarin POS tagging
- Li et al. (2011) propose **joint models** for Chinese POS tagging and dependency parsing to improve both of them

Related work

Multiple resources

- Jiang et al. (2011) propose **a guide feature based method** for Chinese word segmentation and POS tagging
- Sun and Wan (2012) propose **a structured guide feature method** for Chinese lexical processing with heterogeneous annotations
- Qiu et al. (2013) propose joint Chinese word segmentation and POS tagging on heterogeneous annotated Corpora with **multiple task learning**

Motivation

Why use multiple resources?

- Single resource has **two drawbacks**
  - limited scale
  - genre coverage

We also use multiple resources in our work!

Motivation

How we use?

- Jiang(2011)'s guide feature based method
  - **Simple and effective**
  - **But has two shortcomings**
    - Use limited language phenomenon
    - Inefficient because of twice decoding

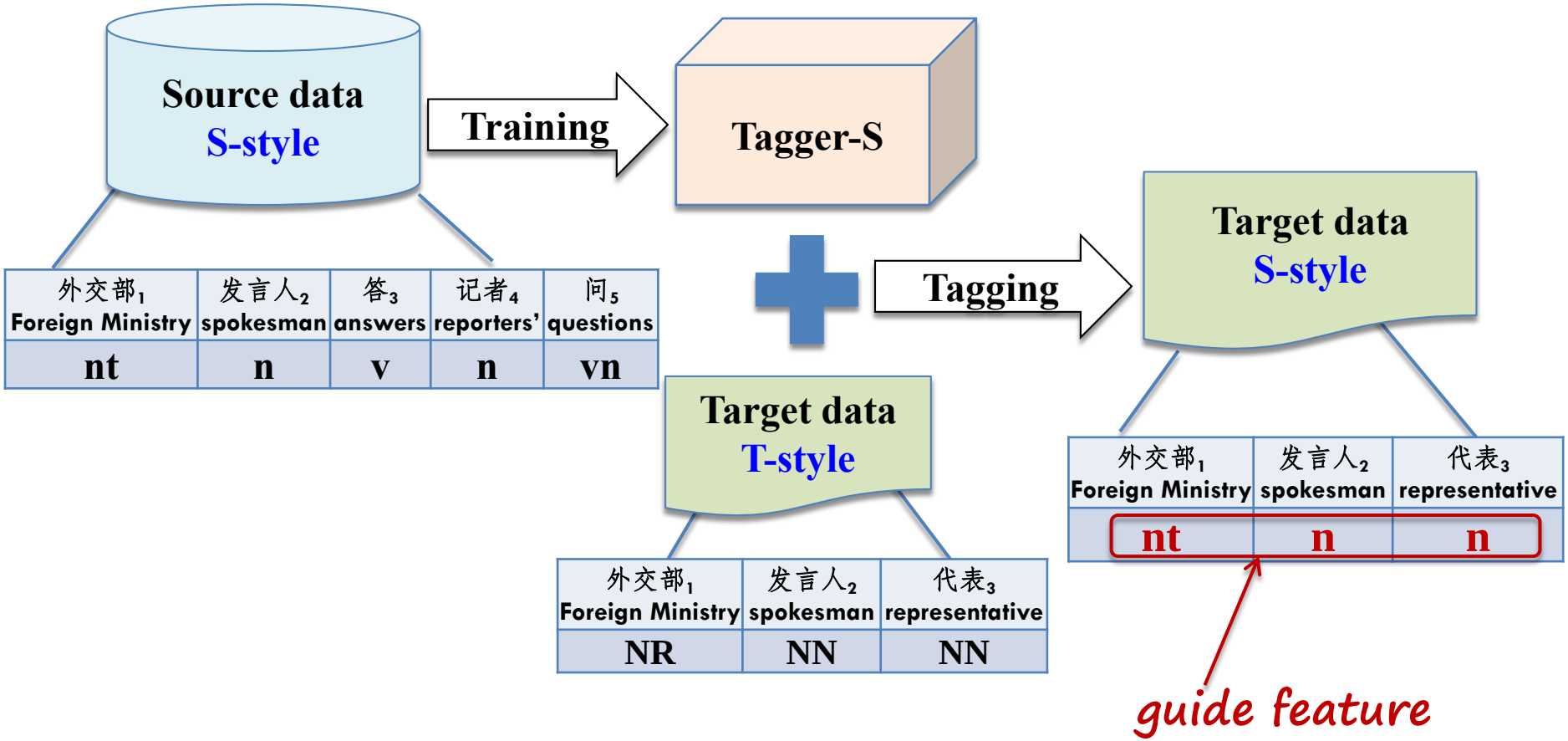
We propose an annotation conversion method!

What is annotation conversion

- **Definition:** convert the source-style annotations into target-style annotations
- **Example**

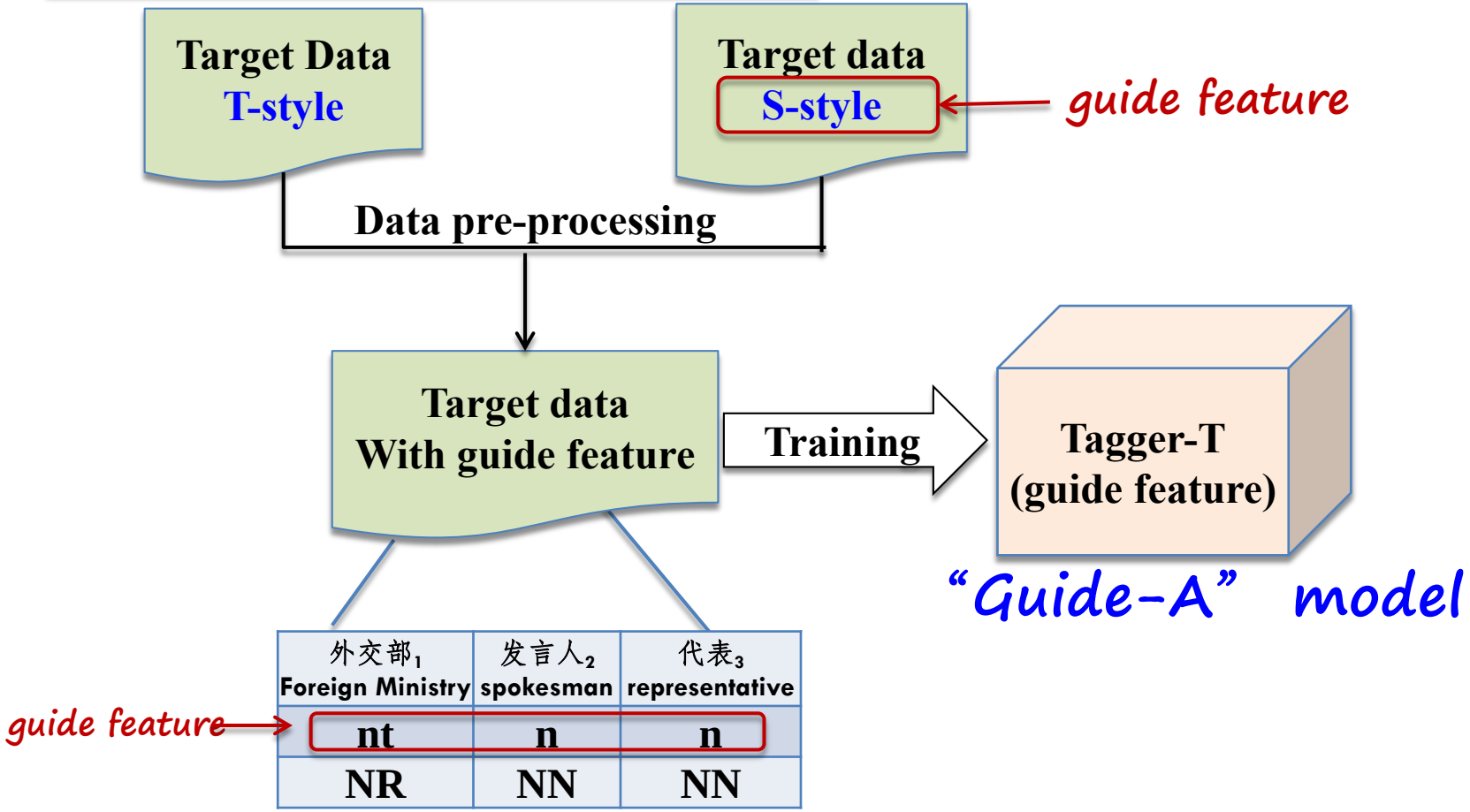
Guide feature based method

Step 1: Get guide feature



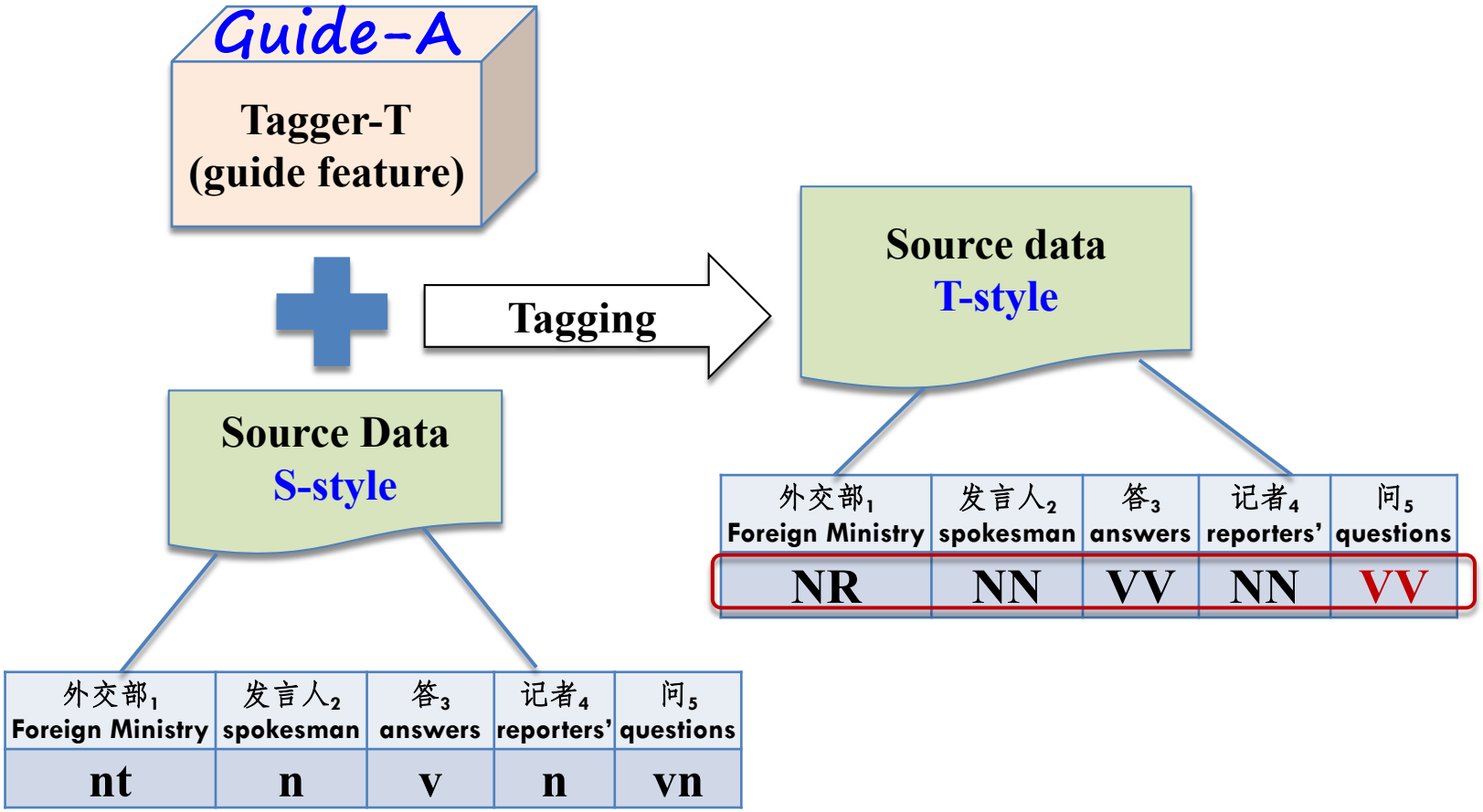
Guide feature based method

Step 2: Train a model



Annotation conversion

Step 3 : Annotation conversation

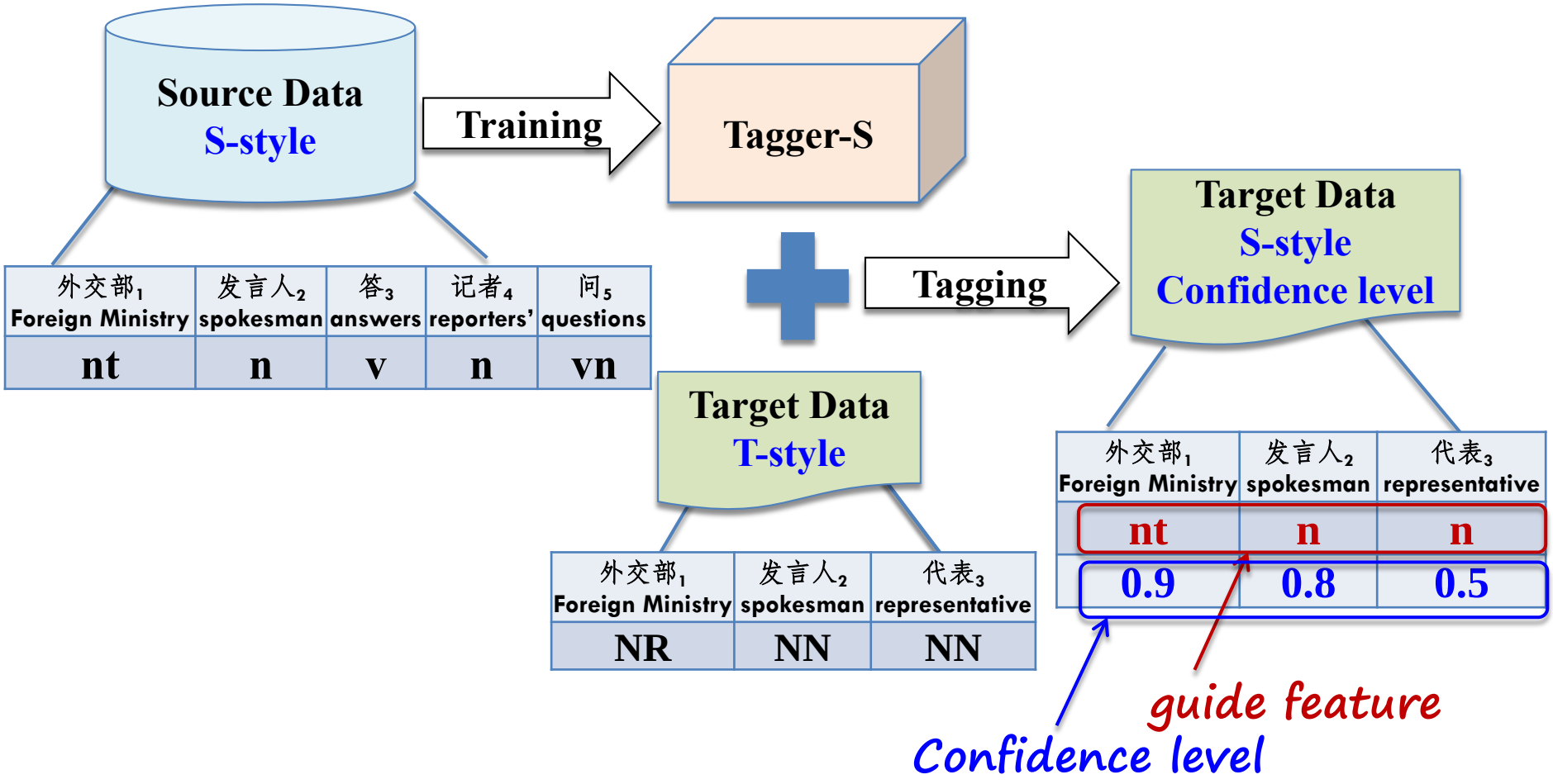


Further

- **Add confidence level to guide feature method**
  - Jiang didn't take confidence level into account in his work
  - Motivated by the intuition that distinguishing the two features can help to improve our conversion performance.
  - Use guide feature's **marginal probability** as its confidence level

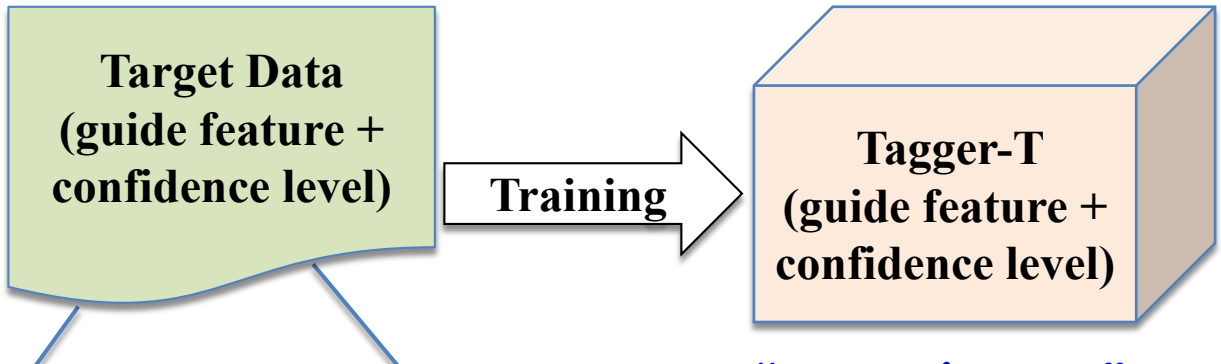
Guide feature based method

Step 1: Get guide feature + Confidence level



Guide feature based method

Step 2: Train a model



"Guide-B" model

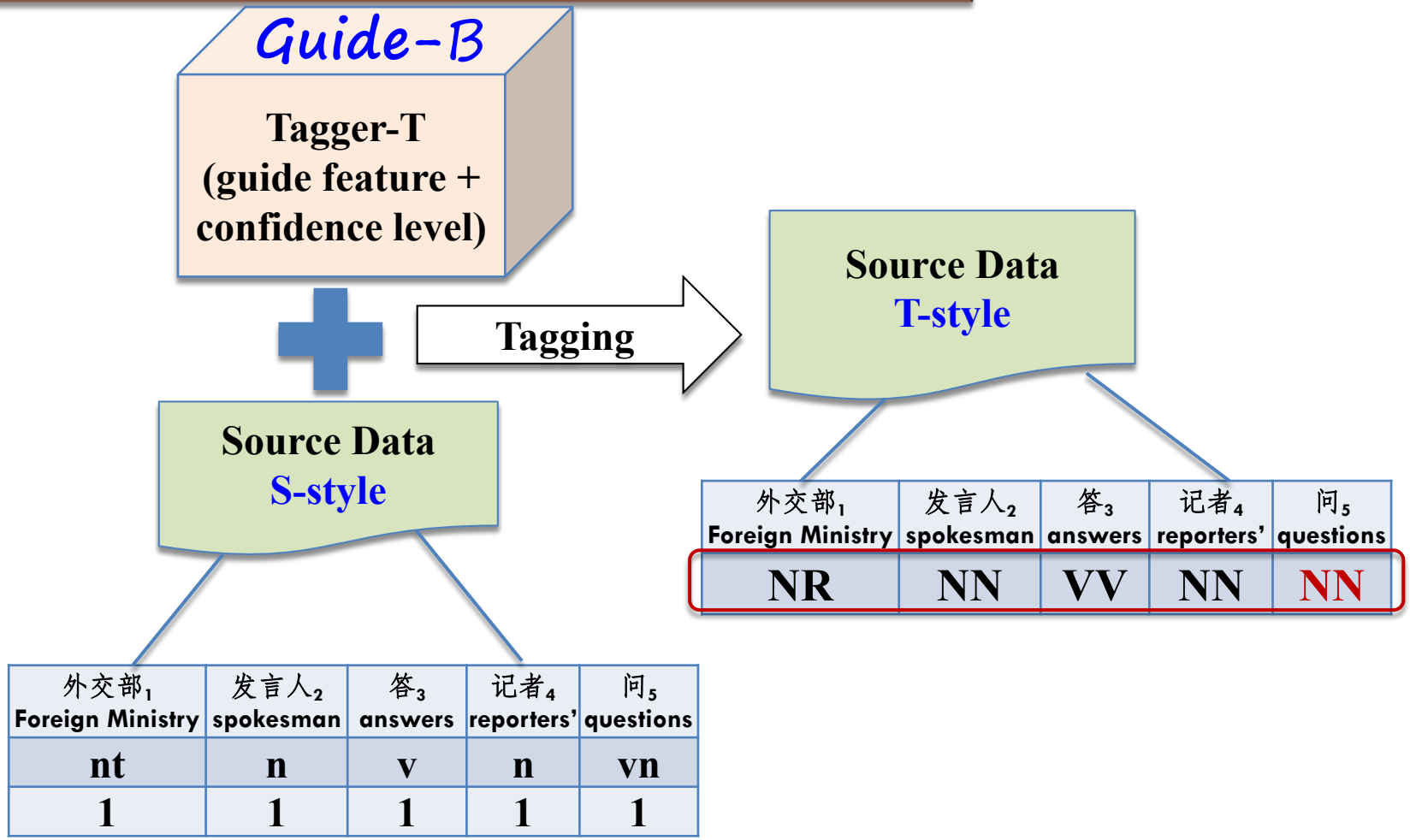
外交部$_1$ Foreign Ministry	发言人$_2$ spokesman	代表$_3$ representative
nt	n	n
0.9	0.8	0.5
NR	NN	NN

guide feature

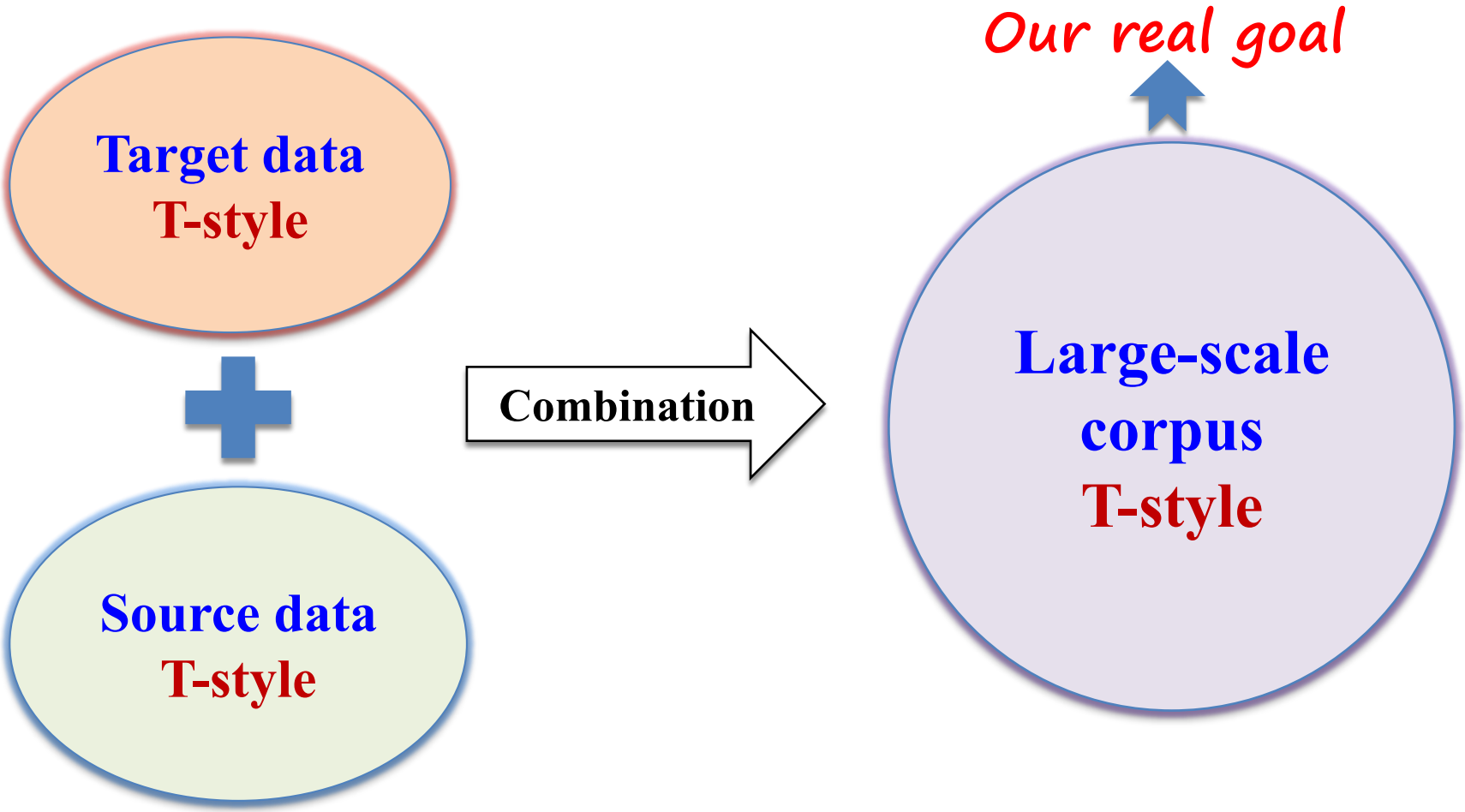
confidence level

Annotation conversation

Step 3 : Annotation conversation

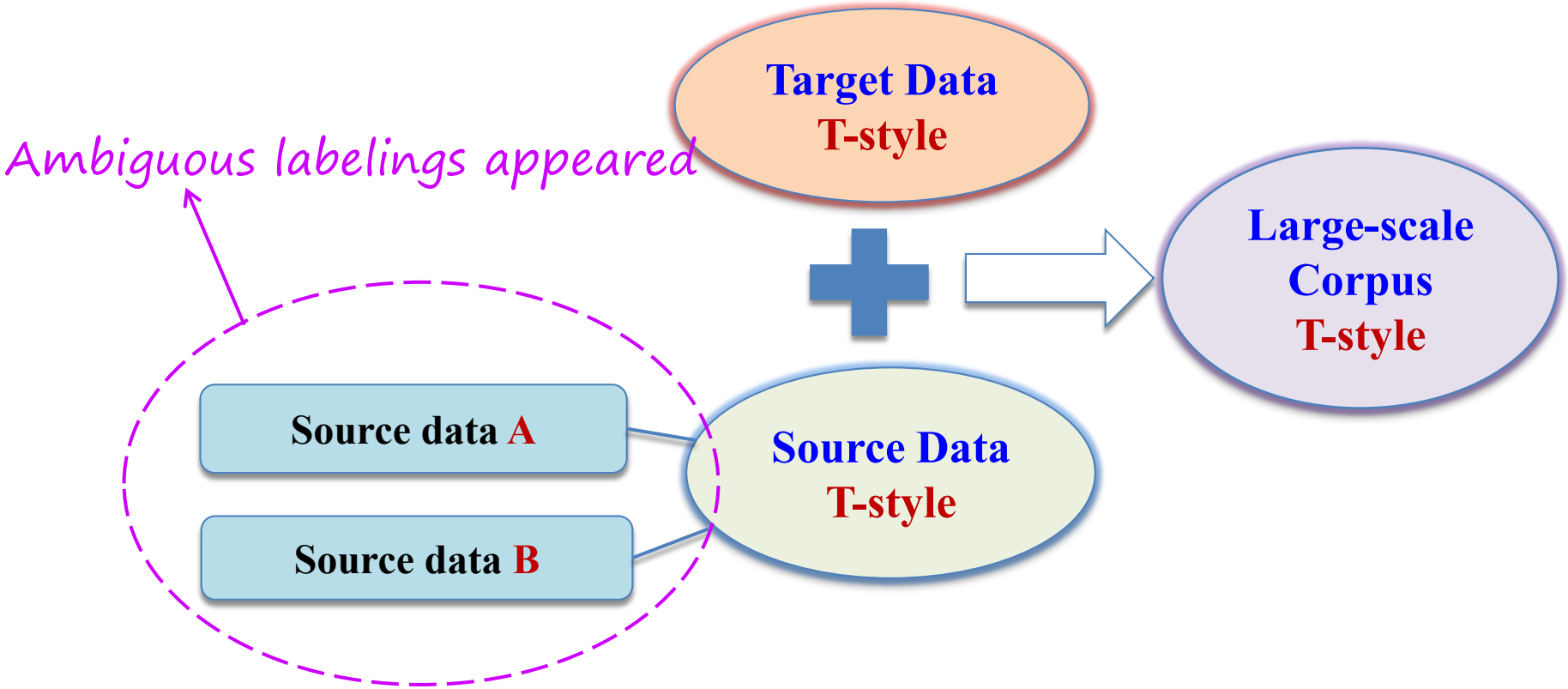


Further

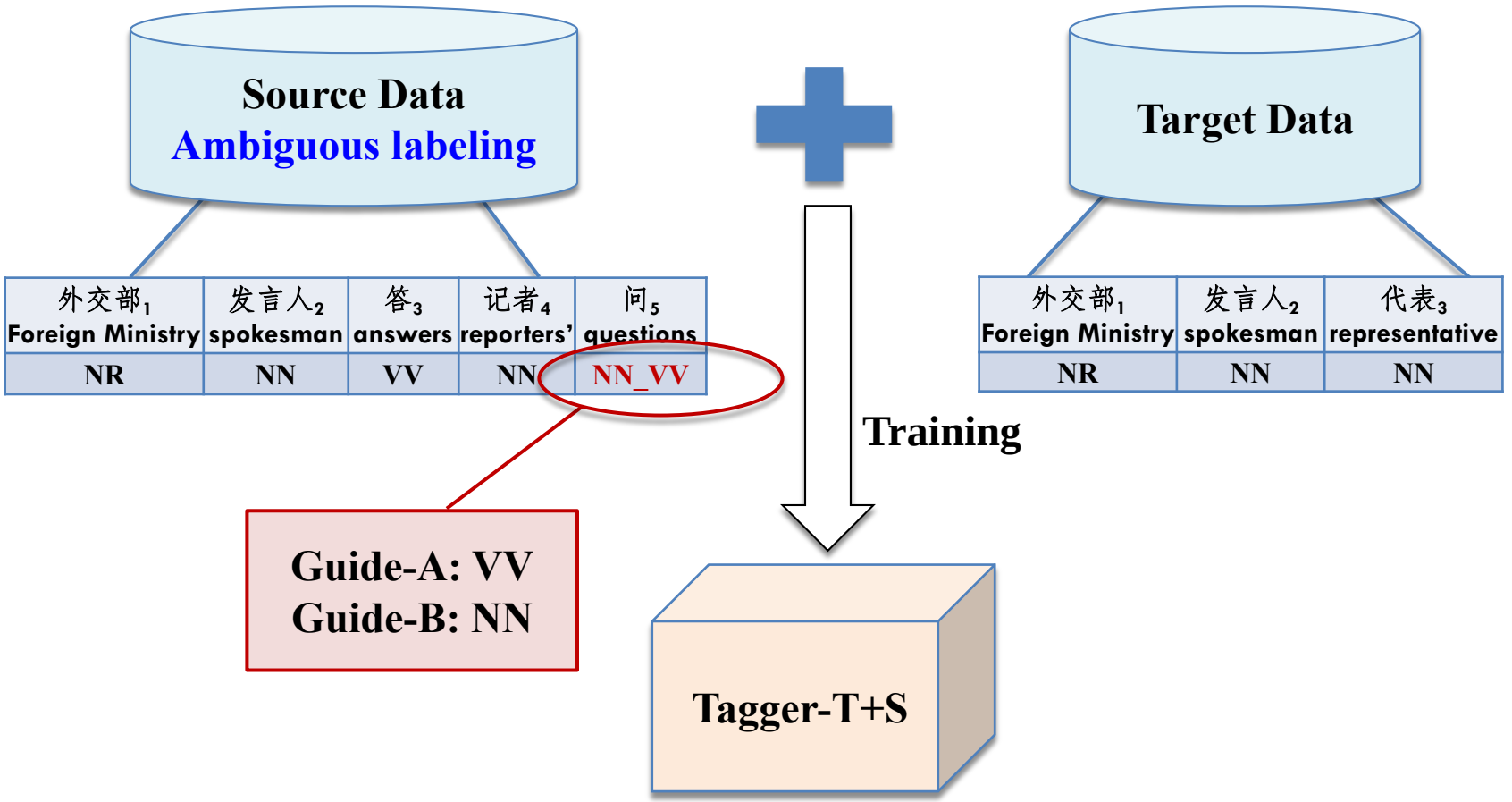

Our real goal
Target data
T-style
Source data
T-style
Combination
Large-scale
corpus
T-style

Further

- **Two versions of target-style annotations**
  - Both of them are got by automatically POS tagging
  - Exist incorrect target-style annotations inevitably



Ambiguous labelings



Extension of CRF

- **The traditional objective is like this:**

  Maximize the probability of gold-standard POS sequence of training data

- **In our case:**

  Maximize the sum probability of all the POS sequence **in the ambiguous labelings**

Experiments

- Corpus statistics
  - Penn Chinese Treebank 5.1 (CTB5)
  - Peking University's People's Daily (PD)

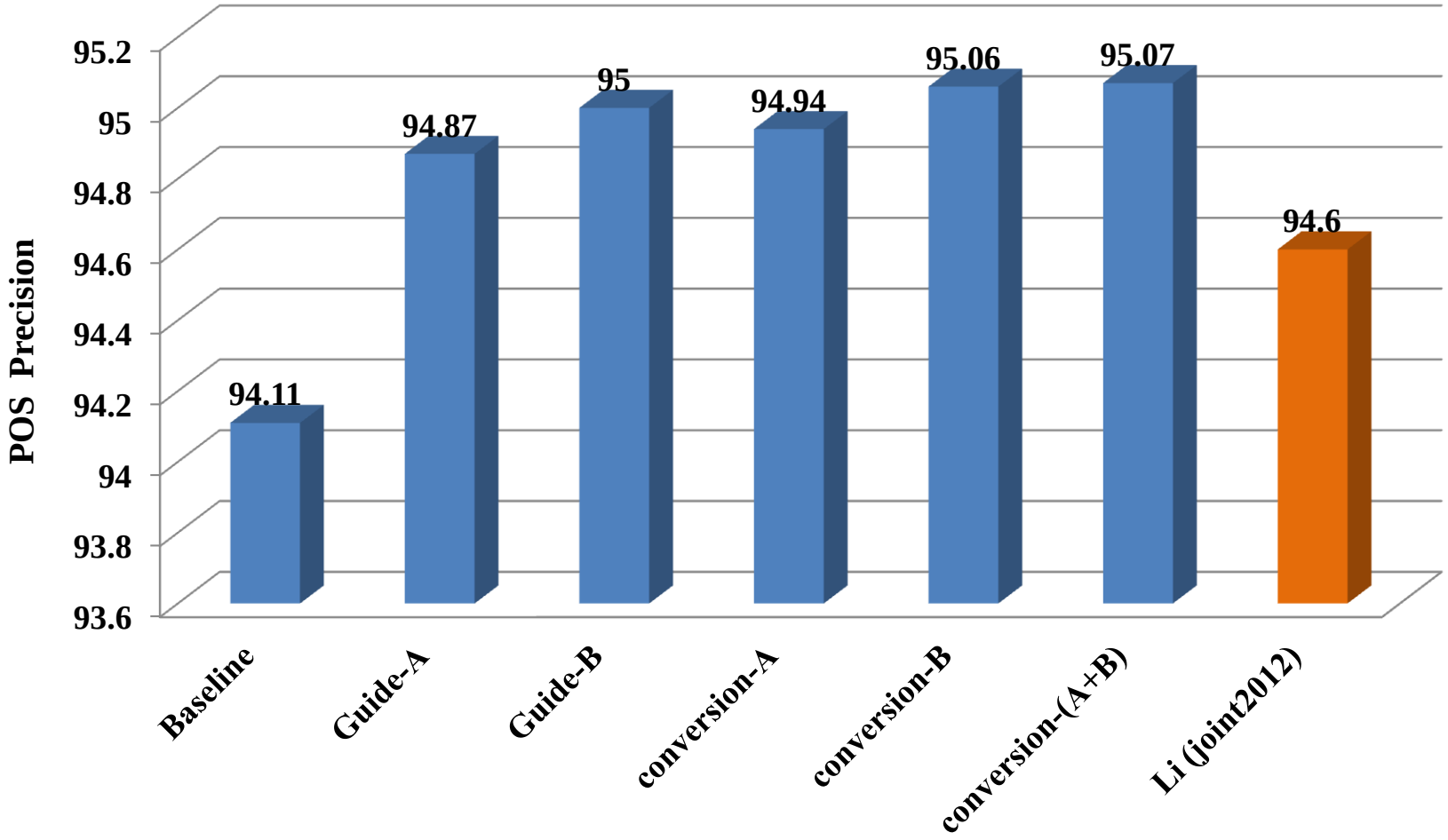
Corpus	Category	Sentences	words
CTB5 (Target data)	Train	16,091	437,991
	Dev	803	20,454
	Test	1,910	50,319
PD (Source data)	Train	291,722	6,843,978
	Dev	5,000	116,887
	Test	10,000	236,528

Experiments

Experiments	Training set
Baseline	CTB5
Guide feature based method (guide feature) [Guide-A]	CTB5+guide feature
Guide feature based method (guide feature + confidence level) [Guide-B]	CTB5+guide feature+confidence level
Annotation conversion [conversion-A]	CTB5 & PD-A
Annotation conversion [conversion-B]	CTB5 & PD-B
Annotation conversion (Ambiguous labelings) [conversion-A+B]	CTB5 & PD-A+B

Experiments

- Experiments on the test set of CTB5



Conclusions

- We proposes **two novel strategies** to advance state-of-the-art methods on multiple resource exploitation.
  - Use reliability information of guide feature and confidence level Use ambiguous labelings to improve the performance of POS tagging

The end

Thank You

Q and A