Semantic Role Labeling for Learner Chinese: the Importance of Syntactic Analysis and L2-L1 Parallel Data

Zi Lin, Yuguang Duan, Yuanyuan Zhao, Weiwei Sun, Xiaojun Wan

Peking University

{*zi.lin, ariaduan, zhao_yy, ws, wanxiaojun*}*@pku.edu.cn*

October 25, 2018

北京大學
PEKING UNIVERSITY

Overview

Background

Data Set

Robustness of L1-annotation-trained SRL Systems

Analysis

Improving SRL Systems with L2-L1 Parallel Data

Outline

Background

Data Set

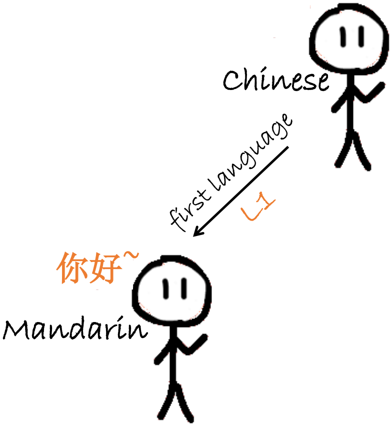
Robustness of L1-annotation-trained SRL Systems

Analysis

Improving SRL Systems with L2-L1 Parallel Data

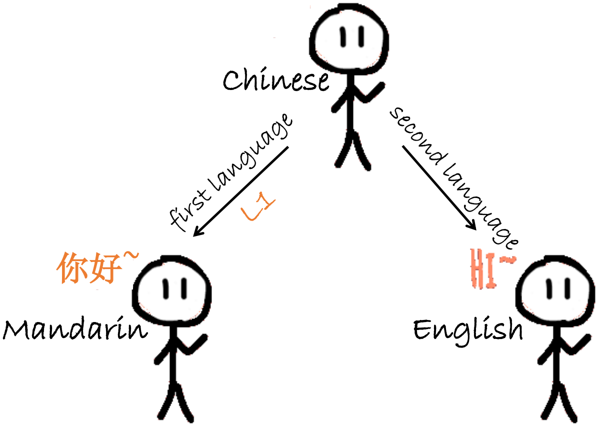
What is interlanguage?

A second language (or L2) which preserves some features of their first language (or L1).



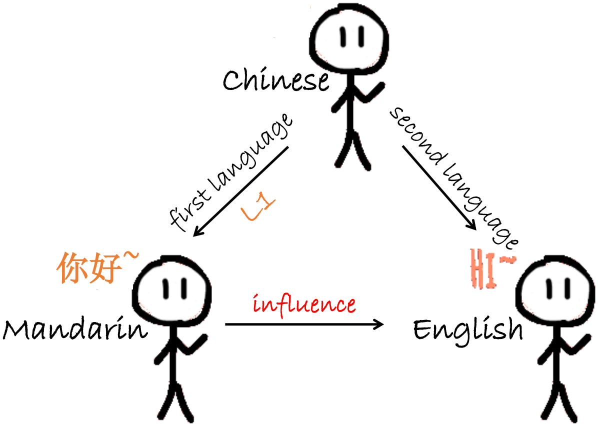
What is interlanguage?

A second language (or L2) which preserves some features of their first language (or L1).

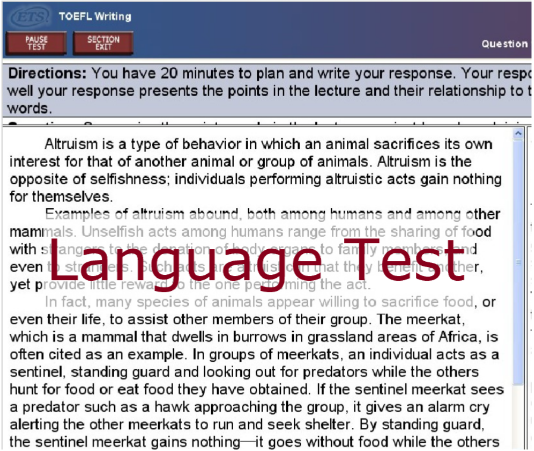


What is interlanguage?

A second language (or L2) which preserves some features of their first language (or L1).



Interlanguage is everywhere...



Directions: You have 20 minutes to plan and write your response. Your respo
well your response presents the points in the lecture and their relationship to t
words.

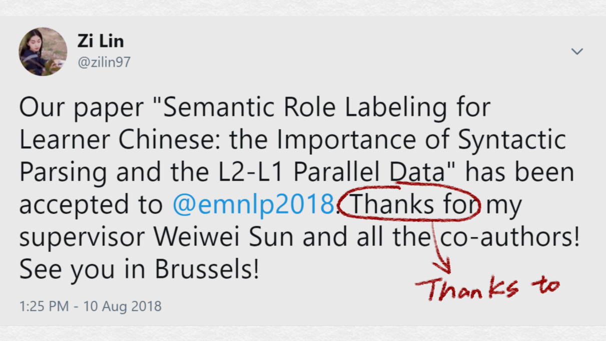
Altruism is a type of behavior in which an animal sacrifices its own interest for that of another animal or group of animals. Altruism is the opposite of selfishness; individuals performing altruistic acts gain nothing for themselves.

Examples of altruism abound, both among humans and among other mammals. Unselfish acts among humans range from the sharing of food with strangers to the donation of body organs to family members, and even to strangers. Such acts are altruistic in that they benefit another, yet provide little reward to the one performing the act.

In fact, many species of animals appear willing to sacrifice food, or even their life, to assist other members of their group. The meerkat, which is a mammal that dwells in burrows in grassland areas of Africa, is often cited as an example. In groups of meerkats, an individual acts as a sentinel, standing guard and looking out for predators while the others hunt for food or eat food they have obtained. If the sentinel meerkat sees a predator such as a hawk approaching the group, it gives an alarm cry alerting the other meerkats to run and seek shelter. By standing guard, the sentinel meerkat gains nothing—it goes without food while the others

Language Test

Interlanguage is everywhere...



Zi Lin
@zilin97

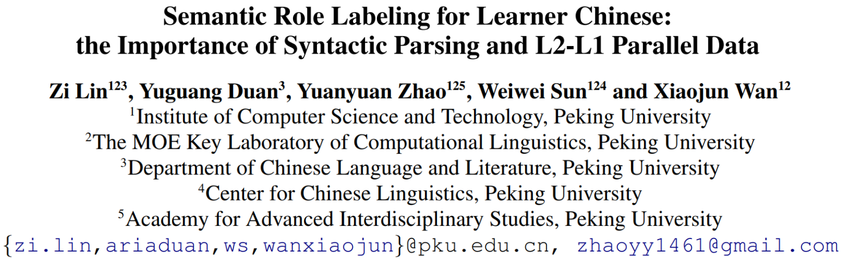
Our paper "Semantic Role Labeling for Learner Chinese: the Importance of Syntactic Parsing and the L2-L1 Parallel Data" has been accepted to @emnlp2018. Thanks for my supervisor Weiwei Sun and all the co-authors! See you in Brussels!

1:25 PM - 10 Aug 2018

Thanks to

Social Network

Interlanguage is everywhere...



Semantic Role Labeling for Learner Chinese: the Importance of Syntactic Parsing and L2-L1 Parallel Data

Zi Lin[1,2,3], Yuguang Duan[3], Yuanyuan Zhao[1,2,5], Weiwei Sun[1,2,4] and Xiaojun Wan[1,2]
[1]Institute of Computer Science and Technology, Peking University
[2]The MOE Key Laboratory of Computational Linguistics, Peking University
[3]Department of Chinese Language and Literature, Peking University
[4]Center for Chinese Linguistics, Peking University
[5]Academy for Advanced Interdisciplinary Studies, Peking University
{zi.lin,ariaduan,ws,wanxiaojun}@pku.edu.cn, zhaoyy1461@gmail.com

And perhaps your paper...

北京大学 PEKING UNIVERSITY

Outline

Background

Data Set

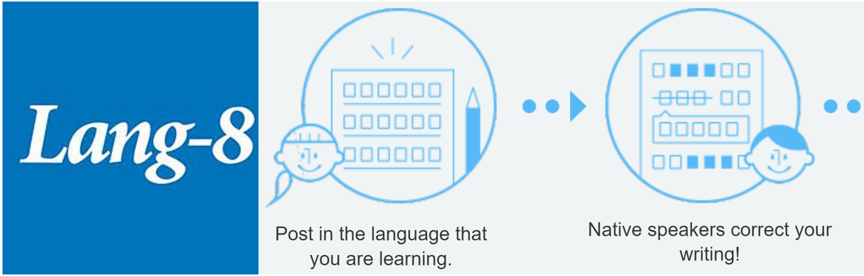
Robustness of L1-annotation-trained SRL Systems

Analysis

Improving SRL Systems with L2-L1 Parallel Data

L2-L1 Parallel Data

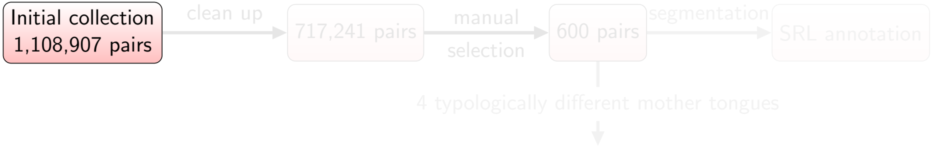
Collect a large dataset of L2-L1 parallel texts of **Mandarin** by exploring "language exchange" social networking services – lang-8[1].



Lang-8
Post in the language that you are learning.
Native speakers correct your writing!

[1] http://lang-8.com/

Data for SRL annotation



Initial collection 1,108,907 pairs

clean up

717,241 pairs

manual selection

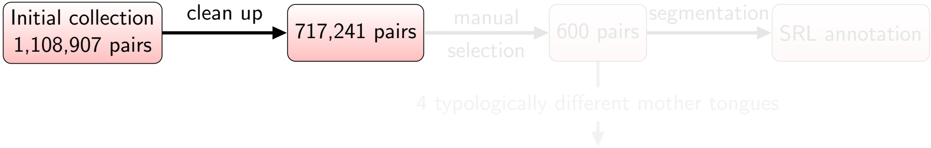
600 pairs

segmentation

SRL annotation

4 typologically different mother tongues

Data for SRL annotation



Initial collection
1,108,907 pairs

clean up

717,241 pairs

manual
selection

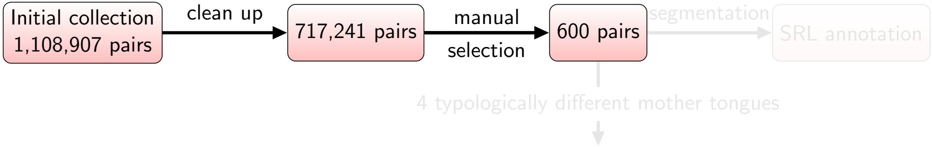
600 pairs

segmentation

SRL annotation

4 typologically different mother tongues

Data for SRL annotation



Initial collection 1,108,907 pairs

clean up

717,241 pairs

manual selection

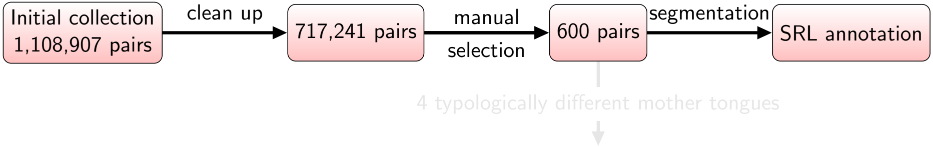
600 pairs

segmentation

SRL annotation

4 typologically different mother tongues

Data for SRL annotation



Initial collection 1,108,907 pairs

clean up

717,241 pairs

manual selection

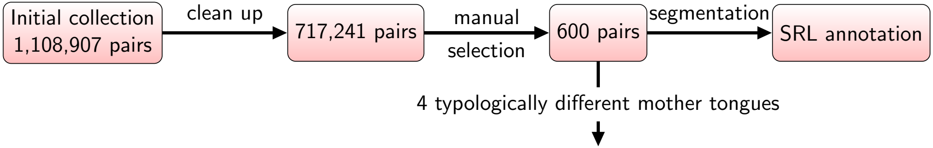
600 pairs

segmentation

SRL annotation

4 typologically different mother tongues

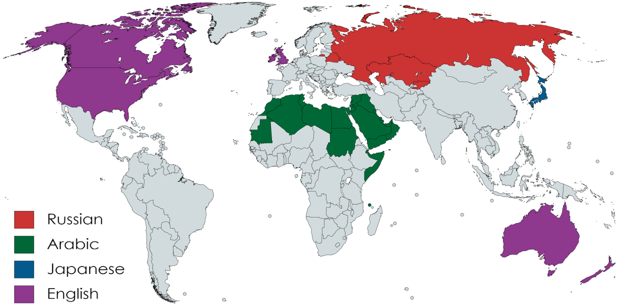
Data for SRL annotation



Initial collection 1,108,907 pairs → clean up → 717,241 pairs → manual selection → 600 pairs → segmentation → SRL annotation

600 pairs → 4 typologically different mother tongues ↓

Language	**Family**
Chinese	Sino-Tibetan
Russian	Slavic
Arabic	Semitic
Japanese	Unknown
English	Germanic



Two Questions

1. Can human understand interlanguage robustly?
2. Can automatic system produce high-quality semantic structures?

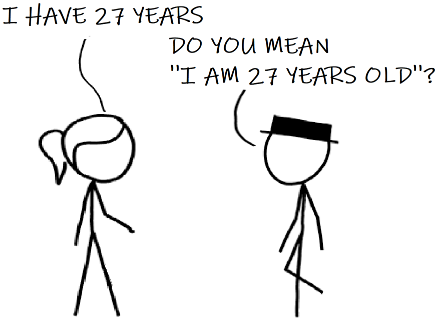
Can human understand interlanguage robustly?

- ☹ It is difficult to define the syntactic formulism of learner language.
- ☺ But sometimes we can understand what they mean...

Why not Semantics?

Can human understand interlanguage robustly?

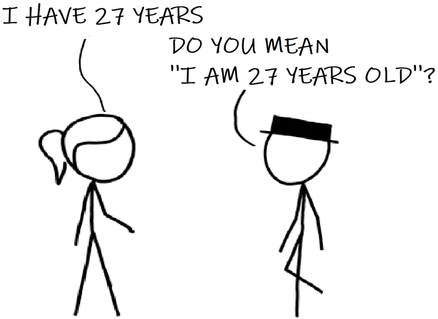
- ☹ It is difficult to define the syntactic formulism of learner language.
- ☺ But sometimes we can understand what they mean...



Why not Semantics?

Can human understand interlanguage robustly?

- ☹ It is difficult to define the syntactic formulism of learner language.
- ☺ But sometimes we can understand what they mean...

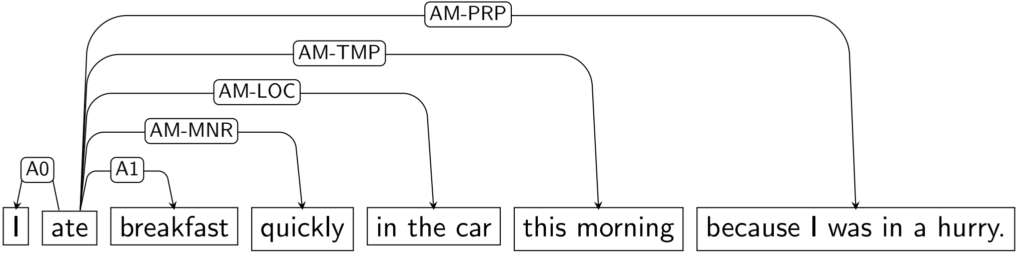


Why not Semantics?

Semantic Role Labeling

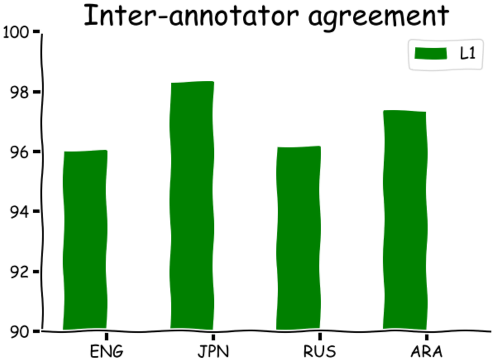
Argument (AN): Who did what to whom?

Adjunct (AM): When, where, why and how?



Inter-annotator agreement

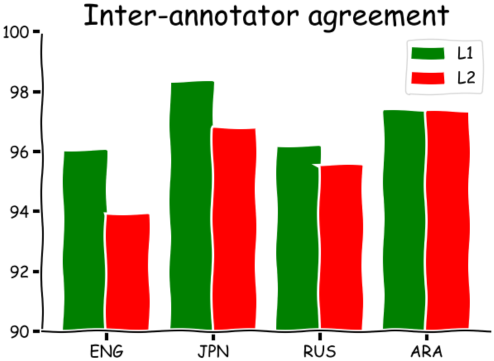
- **Annotator:** two Linguistic students
- **The first 50-sentence trial set:** adapting and refining CPB secification
- **The rest 100-sentence set:** reporting the inter-annotator agreement



北京大學
PEKING UNIVERSITY

Inter-annotator agreement

- **Annotator:** two Linguistic students
- **The first 50-sentence trial set:** adapting and refining CPB secification
- **The rest 100-sentence set:** reporting the inter-annotator agreement



Two Questions

1. Can human understand interlanguage robustly?

2. Can automatic system produce high-quality semantic structures?

Outline

Background

Data Set

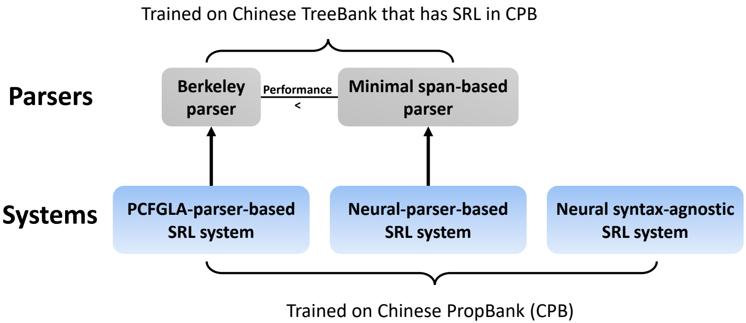
Robustness of L1-annotation-trained SRL Systems

Analysis

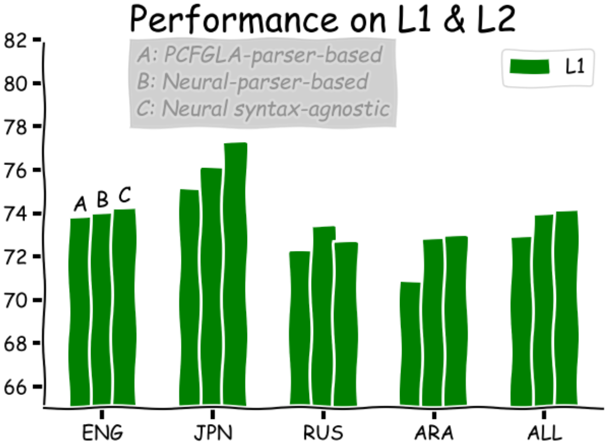
Improving SRL Systems with L2-L1 Parallel Data

Three SRL systems

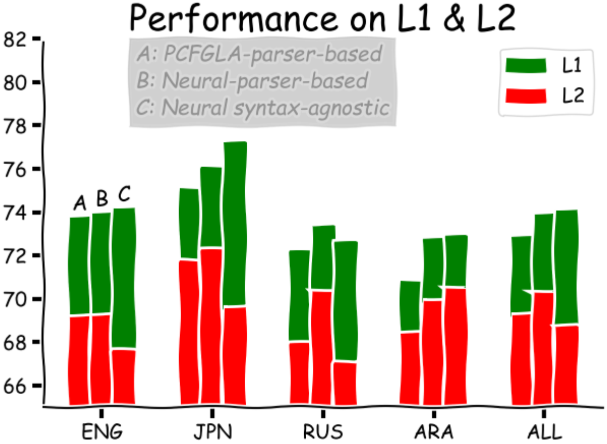
- The Necessity of Parsing for Predicate Argument Recognition. (2002). Gildea and Palmer.
- Semantic Role Labeling Using Different Syntactic Views (2005). Pradhan et al.
- Syntax for Semantic Role Labeling, To Be, Or Not To Be. (2018). He et al.
- Linguistically-Informed Self-Attention for Semantic Role Labeling. (2018). Strubell et al. EMNLP 2018 Best Paper


Trained on Chinese TreeBank that has SRL in CPB
Parsers
Berkeley parser
Performance
<
Minimal span-based parser
Systems
PCFGLA-parser-based SRL system
Neural-parser-based SRL system
Neural syntax-agnostic SRL system
Trained on Chinese PropBank (CPB)

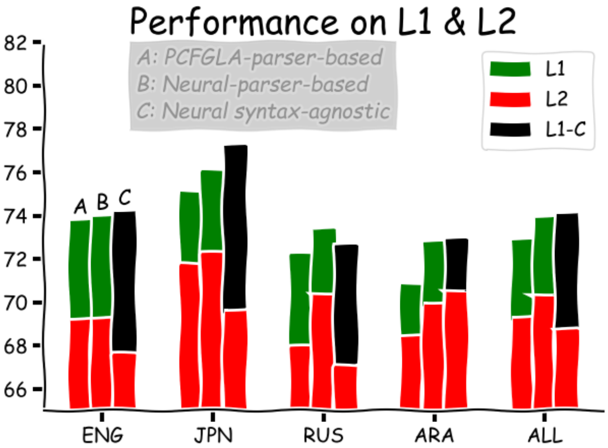
Results


Performance on L1 & L2
A: PCFGLA-parser-based
B: Neural-parser-based
C: Neural syntax-agnostic
L1
82
80
78
76
74
72
70
68
66
A
B
C
ENG
JPN
RUS
ARA
ALL

Results

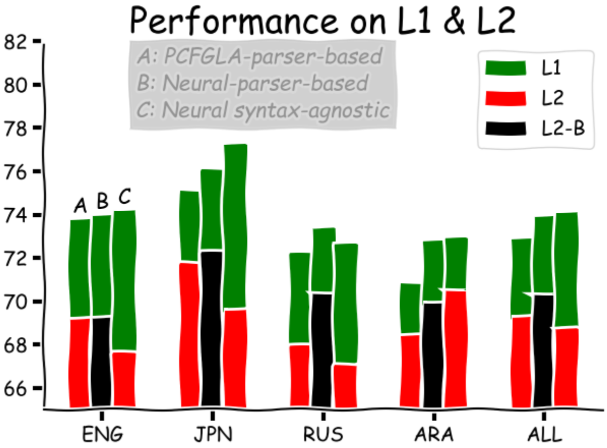

Performance on L1 & L2
A: PCFGLA-parser-based
B: Neural-parser-based
C: Neural syntax-agnostic
🟩 L1
🟥 L2
82
80
78
76
74
72
70
68
66
A
B
C
ENG
JPN
RUS
ARA
ALL

Findings



The syntax-based systems are more robust when handling learner texts.

Findings



The better the parsing results we get, the better the performance on L2 we achieve.

Outline

Background

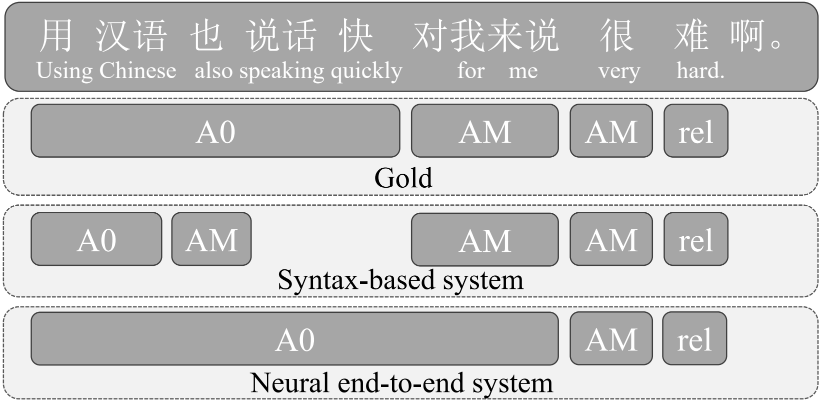
Data Set

Robustness of L1-annotation-trained SRL Systems

Analysis

Improving SRL Systems with L2-L1 Parallel Data

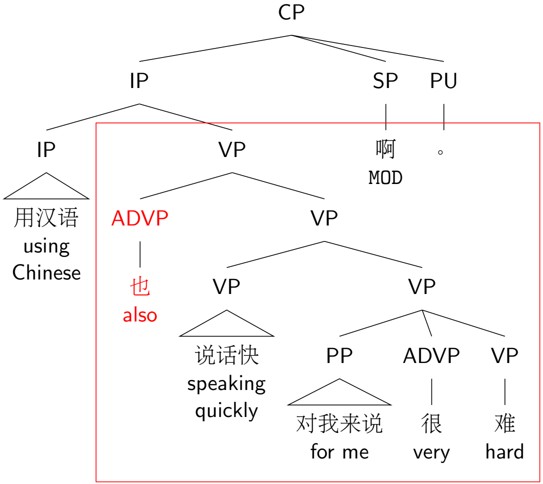
Why syntactic analysis is important?



用 汉语 也 说话 快 对我来说 很 难 啊。
Using Chinese also speaking quickly for me very hard.
A0
AM
AM
rel
Gold
A0
AM
AM
AM
rel
Syntax-based system
A0
AM
rel
Neural end-to-end system

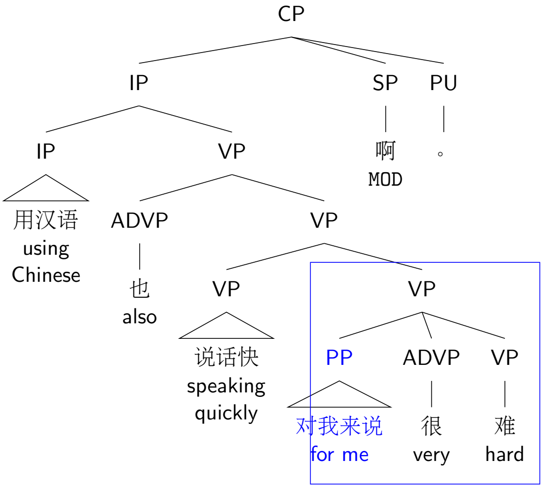
Using Chinese and also speaking quickly is very hard for me

Why syntactic analysis is important?



- Though the whole structure is ill-formed

Why syntactic analysis is important?



- Partial of the sentence can be well-formed.

A new Questions

1. Can human understand interlanguage robustly?
2. Can automatic system produce high-quality semantic structures?

↓

3. Can we improve the SRL performance on interlanguage?

Outline

Background

Data Set

Robustness of L1-annotation-trained SRL Systems

Analysis

Improving SRL Systems with L2-L1 Parallel Data

Leveraging L2-L1 Parallel Data

☺	我	喜欢	做	中国菜
	I	like	cooking	Chinese food

☺	我	喜欢	做饭
	I	like	cooking meal

☹	我	喜欢	做饭	中国菜
	I	like	cook-meal	Chinese food

Leveraging L2-L1 Parallel Data

☺	我	喜欢	做	中国菜
	I	like	cooking	Chinese food

☺	我	喜欢	做饭
	I	like	cooking meal

☹	我	喜欢	做饭	中国菜
	I	like	cook-meal	Chinese food

Leveraging L2-L1 Parallel Data

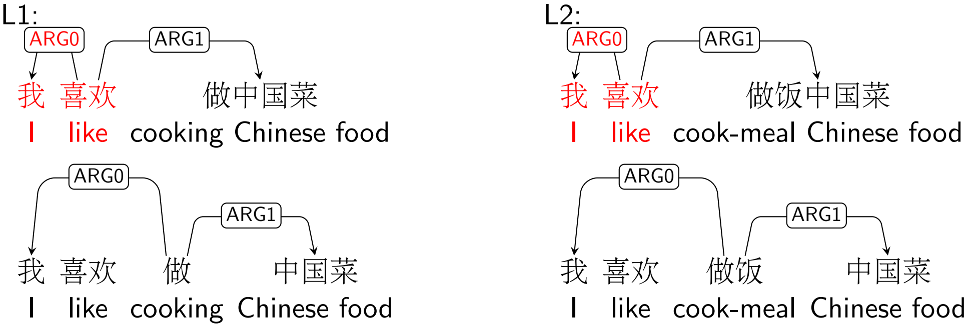
☺	我	喜欢	做	中国菜
	I	like	cooking	Chinese food

☺	我	喜欢	做饭
	I	like	cooking meal

☹	我	喜欢	做饭	中国菜
	I	like	cook-meal	Chinese food

Leveraging L2-L1 Parallel Data

$\langle predicate, argument, role \rangle$ tuples



L1:

ARG0 ARG1

我 喜欢 做中国菜

I like cooking Chinese food

ARG0 ARG1

我 喜欢 做 中国菜

I like cooking Chinese food

L2:

ARG0 ARG1

我 喜欢 做饭中国菜

I like cook-meal Chinese food

ARG0 ARG1

我 喜欢 做饭 中国菜

I like cook-meal Chinese food

$\#$ of shared tuples $= 1$

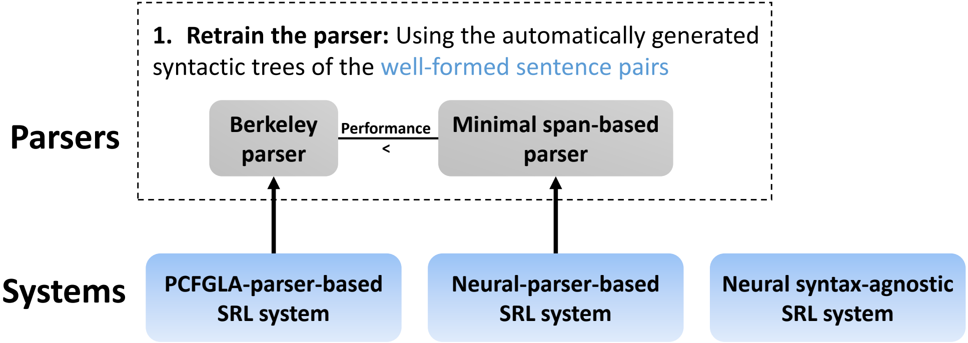
Leveraging L2-L1 Parallel Data

Metric for comparing SRL results

- L2-recall:
  (# of shared tuples) / (# of tuples of the result in L2)
- L1-recall:
  (# of shared tuples) / (# of tuples of the result in L1)

Well-formed sentence pair if both are greater than λ

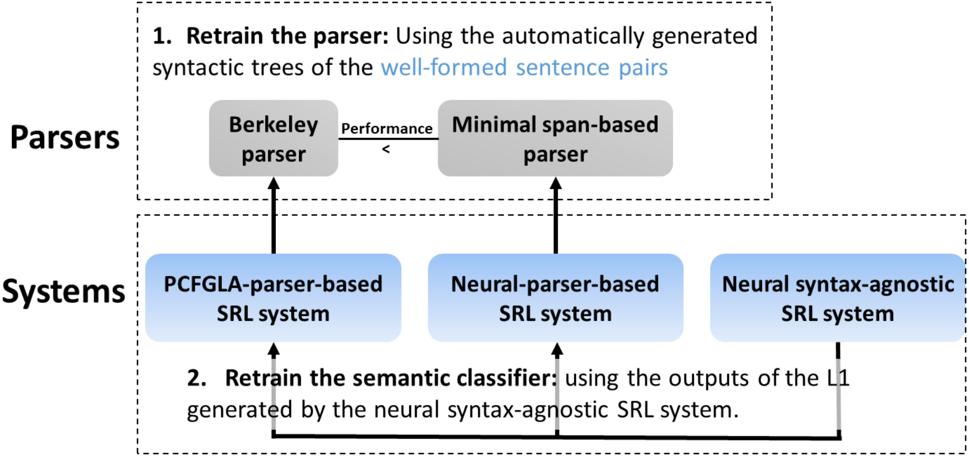
Retraining two essential modules



1. **Retrain the parser:** Using the automatically generated syntactic trees of the well-formed sentence pairs

Parsers
Berkeley parser
Performance
<
Minimal span-based parser
Systems
PCFGLA-parser-based SRL system
Neural-parser-based SRL system
Neural syntax-agnostic SRL system

Retraining two essential modules

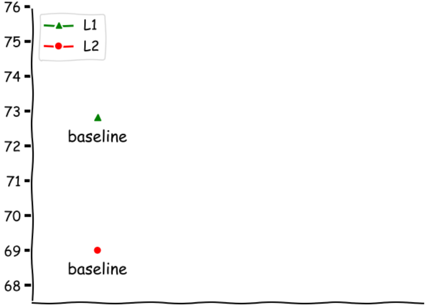


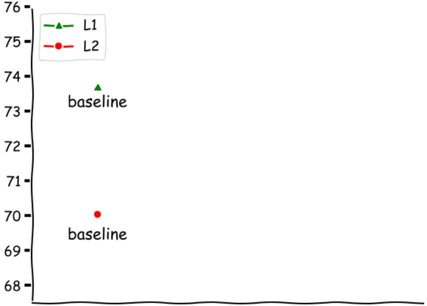
1. Retrain the parser: Using the automatically generated syntactic trees of the well-formed sentence pairs

2. Retrain the semantic classifier: using the outputs of the L1 generated by the neural syntax-agnostic SRL system.

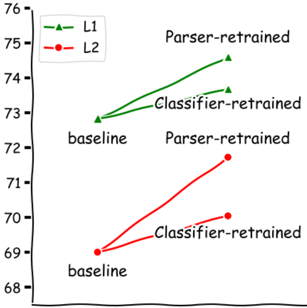
Parsers
Berkeley parser
Performance
<
Minimal span-based parser
Systems
PCFGLA-parser-based SRL system
Neural-parser-based SRL system
Neural syntax-agnostic SRL system

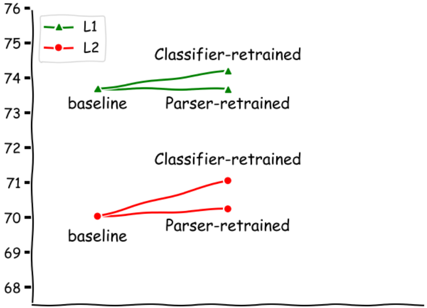
Results


76
75
74
73
72
71
70
69
68
L1
L2
baseline
baseline

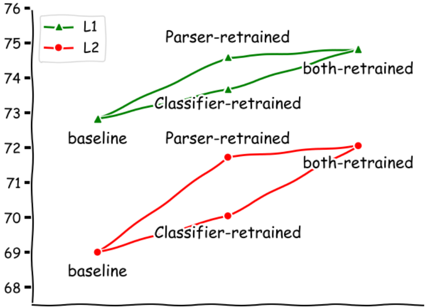

76
75
74
73
72
71
70
69
68
L1
L2
baseline
baseline

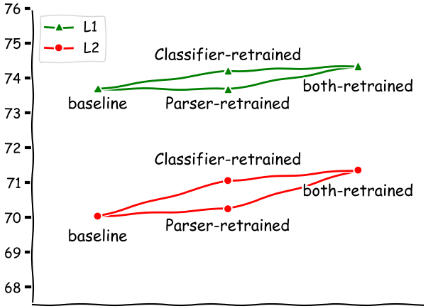
Results


76
75
74
73
72
71
70
69
68
L1
L2
Parser-retrained
Classifier-retrained
baseline
Parser-retrained
Classifier-retrained
baseline


76
75
74
73
72
71
70
69
68
L1
L2
Classifier-retrained
baseline
Parser-retrained
Classifier-retrained
Parser-retrained
baseline

Results


76
75
74
73
72
71
70
69
68
L1
L2
Parser-retrained
both-retrained
Classifier-retrained
baseline
Parser-retrained
both-retrained
Classifier-retrained
baseline


76
75
74
73
72
71
70
69
68
L1
L2
Classifier-retrained
both-retrained
baseline
Parser-retrained
Classifier-retrained
both-retrained
baseline
Parser-retrained

Thanks for your attention!

Zi Lin is planning to apply for PhD program in CS or linguistics this fall. Email me at zi.lin@pku.edu.cn if you are interested!