

# Cross-Lingual Semantic Representation



Weiwei Sun

Institute of Computer Science  
and Technology  
Peking University

July 17, 2019

## Joint work with

---



## Neural string-to-graph parsers are cool!

---

Elementary Dependency Structure	SMATCH	EDM
Factorization-Based	95+	-
Synchronous Hyperedge Replacement Grammar	93+	92+

### Do they touch the upper bound?

Annotator Comparison				
Metric	A vs. B	A vs. C	B vs. C	Average
EDM	94	94	95	94

E. Bender, D. Flickinger, S. Oepen, W. Packard and A. Copestake. 2015. Layers of Interpretation: On Grammar and Compositionality.

## Neural string-to-graph parsers are cool!

---

Elementary Dependency Structure	SMATCH	EDM
Factorization-Based	95+	-
Synchronous Hyperedge Replacement Grammar	93+	92+

### Do they touch the upper bound?

Annotator Comparison				
Metric	A vs. B	A vs. C	B vs. C	Average
EDM	94	94	95	94

E. Bender, D. Flickinger, S. Oepen, W. Packard and A. Copestake. 2015. Layers of Interpretation: On Grammar and Compositionality.

## How good are neural semantic parsers?

---

Hans What are you doing recently?

Weiwei Data-driven graph-based parsing. Deep learning technologies are really cool. I don't know how to improve such parsers. Let me show you some system outputs.

*After examining the very first sentence that contains a parsing error*

Hans I think your parser does the correct thing and the manual annotation is wrong.

### Question

Am I losing my job?

# Cross-lingual parsing

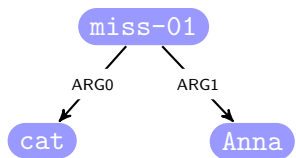
## Multilingual parsing

One single parsing architecture for many languages

- SemEval 2016: Chinese Semantic Dependency Parsing
- SemEval 2019: Cross-lingual semantic parsing with UCCA
  - \* English, German, French

## Cross-lingual parsing

Mapping a string of  $\mathcal{L}_A$  to a graph of  $\mathcal{L}_B$



EN: *Anna's cat is missing her*

DE: *Anna fehlt ihrem Kater*

# Cross-lingual things

---

## Motivation

- Claim: Don't create annotations for  $\mathcal{L}_A$ .
- Secret: Now the system accuracy is low enough for me to improve.

# Cross-lingual things

---

## Motivation

- Claim: Don't create annotations for  $\mathcal{L}_A$ .
- Secret: Now the system accuracy is low enough for me to improve.





# Cross-lingual things

---

## Motivation

- Claim: Don't create annotations for  $\mathcal{L}_A$ .
- Secret: Now the system accuracy is low enough for me to improve.



## Machine Translation before the deep learning era

- State-of-the-art: string-to-tree, tree-to-tree, tree-to-string, etc.
- What is on the way: string-to-graph, graph-to-graph, graph-to-string, etc.

# This talk

---

## What we have done

- Translating c.a. 3000 sentences in Redwoods to Mandarin Chinese.
- Add some annotation layers, currently GB-style syntactic analysis.

## What confuse us

- It is a good idea to use  $\mathcal{L}_B$ 's semantic graphs represent  $\mathcal{L}_B$ 's meanings?
- If not, what is a plausible way to cross languages?

## What we want to do

- Cross-lingual, shared ontological, compositional semantic construction

# Outline

---

Initialization: Translating Redwoods

Problem: Information-appropriateness

Proposal: Shared-ontological Representation

# Outline

---

Initialization: Translating Redwoods

Problem: Information-appropriateness

Proposal: Shared-ontological Representation

## Translating Redwoods

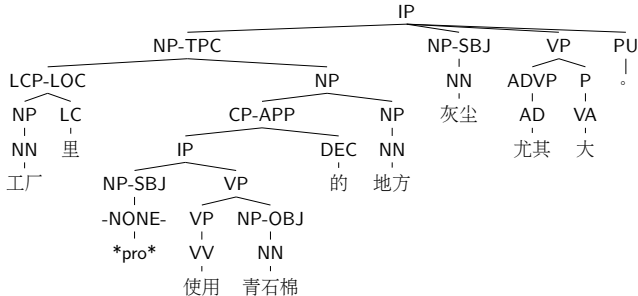
---

- Some sentences are translated by a native speaker.
- Some sentences are first translated by a bilingual and then revised by a native speaker.
  - Mandarin and English-speaking Chinese Singaporeans and Malaysians

<b>Source</b>	<b>Translator</b>	<b>#sentence</b>	<b>#word</b>
wsj	native speaker	1266	31
	bilingual→native speaker	1003	25
	total	2269	29
wescience	bilingual→native speaker	920	21

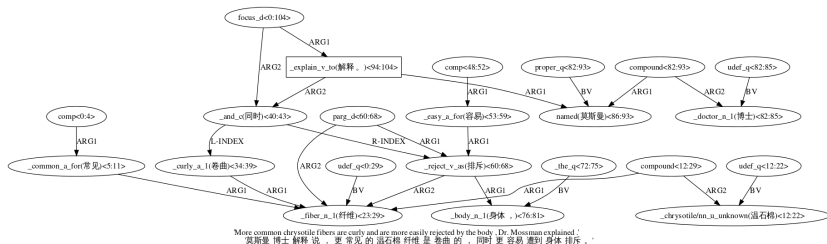
# Annotating translated sentences

- Word alignment (100 sentences)
  - Berkeley aligner
  - Manual correction
- TreeBanking (100 sentences)
  - Following Chinese TreeBank, a PTB-style treebank for Mandarin.



[wsj#20003025]

# A cross-lingual semantic graph



*More common chrysotile fibers are curly and are more easily rejected by the body, Dr. Mossman explained.*

[WSJ, #20003021]

# Outline

---

Initialization: Translating Redwoods

Problem: Information-appropriateness

Proposal: Shared-ontological Representation



# Predication

---

1. shore up a decline [wsj#20012010]  
支撑……减少的局面/situation
2. four-color page [wsj#20012005]  
四色印刷页面
3. was for dinner and dancing [wsj#20010016]  
是……晚餐舞会  
▷coordination  
▷compound
4. from the same period last year [wsj#20011007]  
同比  
▷we have a single word
5. were particularly dusty/ADJ [wsj#20003025]  
灰尘/NOUN尤其大/ADJ

# Predication

---

1. shore up a decline [wsj#20012010]  
支撑……减少的局面/situation
2. four-color page [wsj#20012005]  
四色印刷页面
3. was for dinner and dancing [wsj#20010016]  
是……晚餐舞会  
▷coordination  
▷compound
4. from the same period last year [wsj#20011007]  
同比  
▷we have a single word
5. were particularly dusty/ADJ [wsj#20003025]  
灰尘/NOUN尤其大/ADJ

# Predication

---

1. shore up a decline [wsj#20012010]  
支撑……减少的局面/situation
2. four-color page [wsj#20012005]  
四色印刷页面
3. was for dinner and dancing ▷coordination  
是……晚餐舞会 ▷compound  
[wsj#20010016]
4. from the same period last year [wsj#20011007]  
同比 ▷we have a single word
5. were particularly dusty/ADJ [wsj#20003025]  
灰尘/NOUN尤其大/ADJ

# Predication

---

1. shore up a decline [wsj#20012010]  
支撑……减少的局面/situation
2. four-color page [wsj#20012005]  
四色印刷页面
3. was for dinner and dancing [wsj#20010016]  
是……晚餐舞会  
▷coordination  
▷compound
4. from the same period last year [wsj#20011007]  
同比  
▷we have a single word
5. were particularly dusty/ADJ [wsj#20003025]  
灰尘/NOUN尤其大/ADJ

# Predication

---

1. shore up a decline [wsj#20012010]  
支撑……减少的局面/situation
2. four-color page [wsj#20012005]  
四色印刷页面
3. was for dinner and dancing [wsj#20010016]  
是……晚餐舞会  
▷coordination  
▷compound
4. from the same period last year [wsj#20011007]  
同比  
▷we have a single word
5. were particularly dusty/ADJ [wsj#20003025]  
灰尘/NOUN尤其大/ADJ

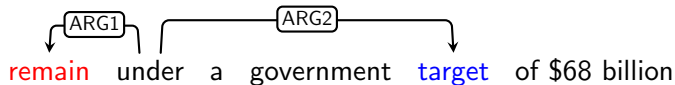
## Predicate–Argument Structure

---

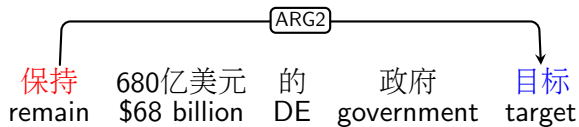
1. remain under a government target of \$68 billion  
保持680亿美元的政府目标 [wsj#20011005]
2. squeezed in a few meetings  
挤出时间开了几个会 [wsj#20010015]
3. mechanically  
用机器 [wsj#20003026]
4. casting a cloud on South Korea 's export-oriented economy  
韩国出口导向型经济笼罩在一片阴云之中 [wsj#20011002]
5. asbestos workers studied in  
研究石棉工人 [wsj#20003017]

# Predicate-Argument Structure

## DeepBank



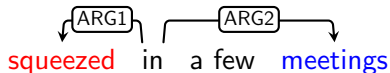
## Mandarin



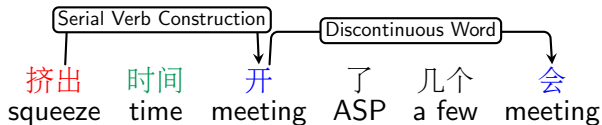
# Predicate-Argument Structure

---

## DeepBank



## Mandarin

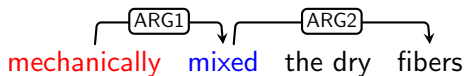




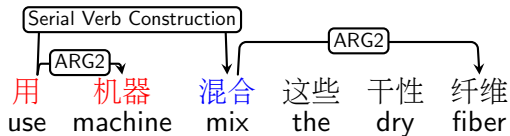
# Predicate-Argument Structure

---

## DeepBank



## Mandarin



# Outline

---

Initialization: Translating Redwoods

Problem: Information-appropriateness

Proposal: Shared-ontological Representation

# Cross-lingual, shared-ontological semantic representation

---

## Annotation tool

<http://59.108.48.37:9014/omg/>

A Glue-like method to semantic composition

# Game over

---

**Q** 生命的意义是什么？

**A** life\_v\_1 →

## Game over

---

**Q** 生命的意义是什么？

**A** [\\_life\\_v\\_1](#) →

Thank You!