Combining corpus statistics and knowledge base to disambiguate and acquire verb frames

學生: 高定慧

指導教授: 張俊盛 博士

系所: 資訊工程學系

時間: 2013.07.09

Abstract

- Verb frames may play a key role in language learning
- Purpose
  - Automatically generate verb frames with semantic labels
- Method
  - Combining corpus statistics and knowledge base
  - Disambiguate semantic labels
- Results
  - We achieved quite satisfied performance comparing to a manually built gold standard

A verb frame

consists of one or more parts which express the requirements for their possible participants.

Tesnière (1953)

Verb frames in different forms

sentence: “I finally abandon the wild idea”

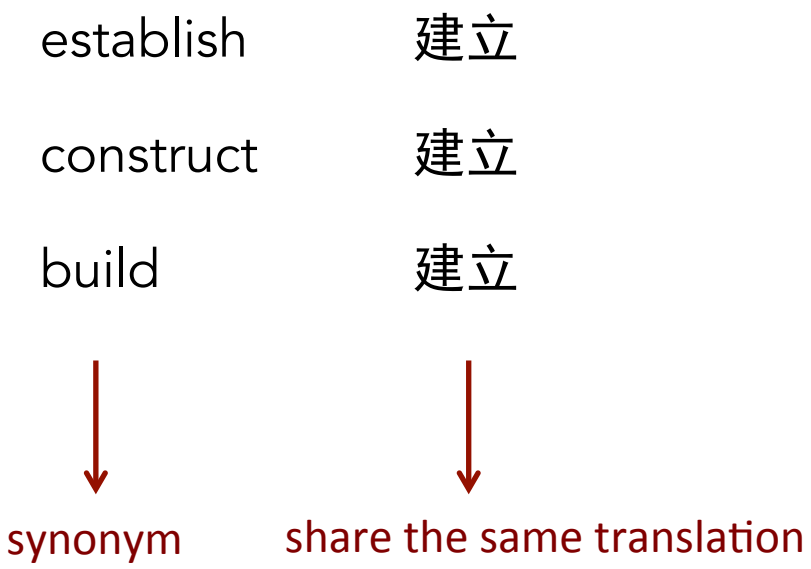
Linggle →	I	abandon	idea
VerbNet →	np	abandon	np
VerbNet →	subject	abandon	object
Dictionary →	somebody	abandon	something
FrameFinder →	person	abandon	cognition
CPA →	person	abandon	plan

Learning verb frames is important

For example, a language learner who is already familiar with:

"members + *establish* + friendship"

Learners tend to rephrase using *similar* words



Unfortunately, this often leads to a word choice error

members establish friendship

* members construct friendship

members build friendship

Dictionaries typically overly simplified semantic labels

1 to build something
construct something from/of/in something

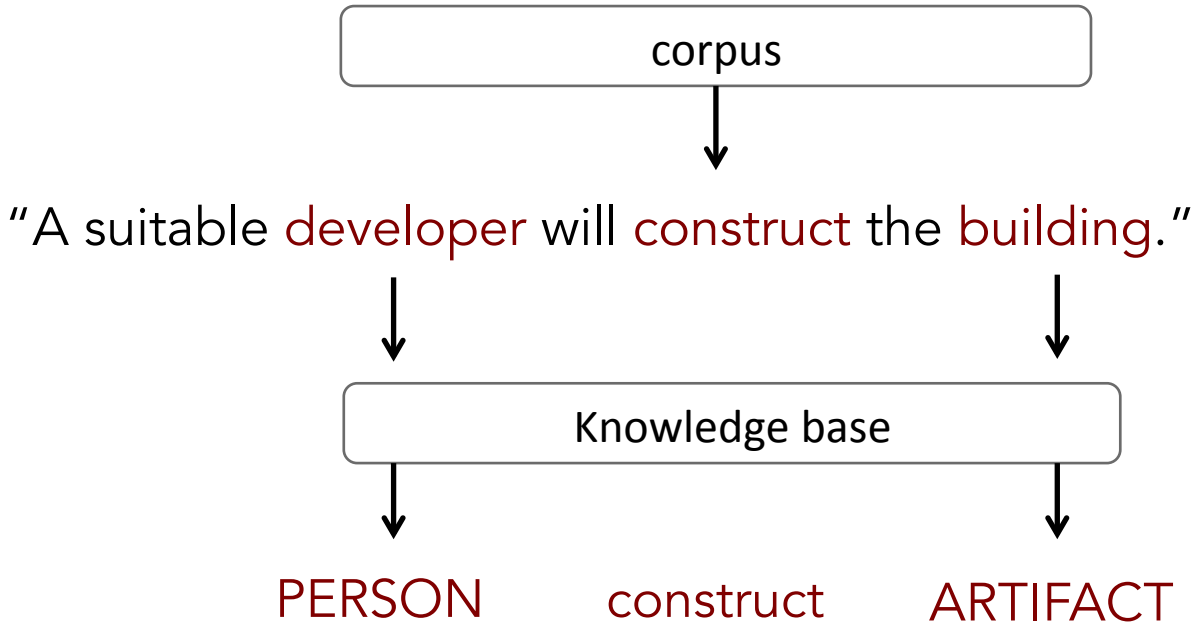
2 to form something

More specific semantic labels help one to use a verb more appropriately

PERSON establish RELATIONSHIP

PERSON construct ARTIFACT

We can easily derive such verb frames from a corpus and a knowledge base

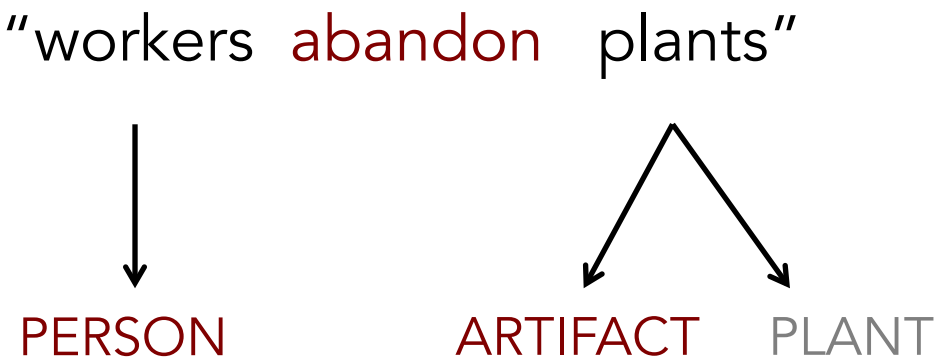


verb construct show 3 instance > 3 pattern 0.0 % search

1	**PERSON construct ARTIFACT**	27	29.35 %
	him construct kiln		Grumpily Dai Huang chose Li Lu , to help him construct a newer , bigg
	you construct shelter		There may have been so much conflict indoors that you had to constru
	I construct bomb		As for Meredith-Lee 's death , if you 're asking me did I construct a bo answer is no .
2	**COGNITION construct ARTIFACT**	18	19.57 %
	intention construct road		Taken in conjunction the written statement and the key diagram indica an A fifty nine relief road , passing to the north of Harrogate and Knar
	techniques construct houses		For the first time , archaeologists have been able to study in detail the medieval builders to construct the typical cob houses for which the We
	example construct building		Would a trust , for example to construct a public building or to set up be enforced as such ?
3	**PERSON construct COGNITION**	13	14.13 %
	I construct tactic		I 've always been too shy to construct a tactic in order to attract wome
	I construct idea		In a split-second I would build on this particle of noise and construct a that could produce such a phenomenon .
	he construct understanding		Unlike the Chicago School , for example , he was not attempting to co how distinct forms of natural area were resulting from people 's searc
4	**PERSON construct COMMUNICATION**	8	8.7 %
	you construct movies		Flip through the wacky but informative manual which casts you as a m that the first three programs let you construct 20-frame icon-sized mo
	we construct database		This we have used to construct and maintain a database of manageme
	he construct discourse		Furthermore , he was attempting to construct a discourse which was ir attack then being launched from inside and outside the discipline .
5	**COGNITION construct COMMUNICATION**	6	6.52 %

Ambiguity existed in a verb frame.

→ *Word Sense Disambiguation*



Many resources provide verb frames information in various forms

Linggle →	I	abandon	idea
VerbNet →	np	abandon	np
VerbNet →	subject	abandon	object
Dictionary →	somebody	abandon	something
FrameFinder →	person	abandon	cognition
CPA →	person	abandon	plan

VerbNet

FRAMES REF KEY

NP V NP

EXAMPLE	"David constructed a house."
SYNTAX	AGENT V RESULT
SEMANTICS	NOT(EXIST(START(E), RESULT)) EXIST(RESULT(E), RESULT) CAUSE(AGENT, E)

NP V NP PP.MATERIAL

EXAMPLE	"David constructed a house out of sticks."
SYNTAX	AGENT V RESULT {FROM OUT_OF} MATERIAL
SEMANTICS	NOT(EXIST(START(E), RESULT)) EXIST(RESULT(E), RESULT) MADE_OF(RESULT(E), RESULT, MATERIAL) CAUSE(AGENT, E)

NP V NP PP.BENEFICIARY

EXAMPLE	"David dug a hole for me."
SYNTAX	AGENT V RESULT {FOR} BENEFICIARY
SEMANTICS	NOT(EXIST(START(E), RESULT)) EXIST(RESULT(E), RESULT) CAUSE(AGENT, E) BENEFIT(E, BENEFICIARY)

NP V NP PP.ATTRIBUTE

EXAMPLE	"They designed the Westinghouse-Mitsubishi venture as a non-equity transaction."
SYNTAX	AGENT V RESULT (AS) ATTRIBUTE
SEMANTICS	NOT(EXIST(START(E), RESULT)) EXIST(RESULT(E), RESULT) CAUSE(E, AGENT)

Kipper-Schuler 2006

[noun] construct [noun] Search Search options

next page

No.	Chunks	Frequency	Examples (from BNC)	Parent (more general versions)
1	[noun sg] **construct a** [noun sg] of	20		
2	[noun] **construct a** [noun sg] of	34		
3	[noun] **be** constructed from [noun]	15		
4	[noun] **construct a** [noun sg]	91		
5	[noun sg] **construct a** [noun sg]	57		
6	[noun] **be** constructed in [noun]	9		
7	[noun] **be** constructed [prep] the [noun]	28		
8	[noun] **be** constructed by [noun]	12		
9	[noun sg] of constructing **a** [noun sg]	16		
10	the [noun sg] of constructing [noun]	7		
11	[noun] constructed by the [noun]	10		
12	[noun] **construct** a new [noun sg]	12		
13	[noun] **be** constructed [prep] the [noun sg]	22		

Wible and Tsao 2010

Corpus Pattern Analysis (CPA)

Human construct Artifact | Building

Human construct Theory | Hypothesis

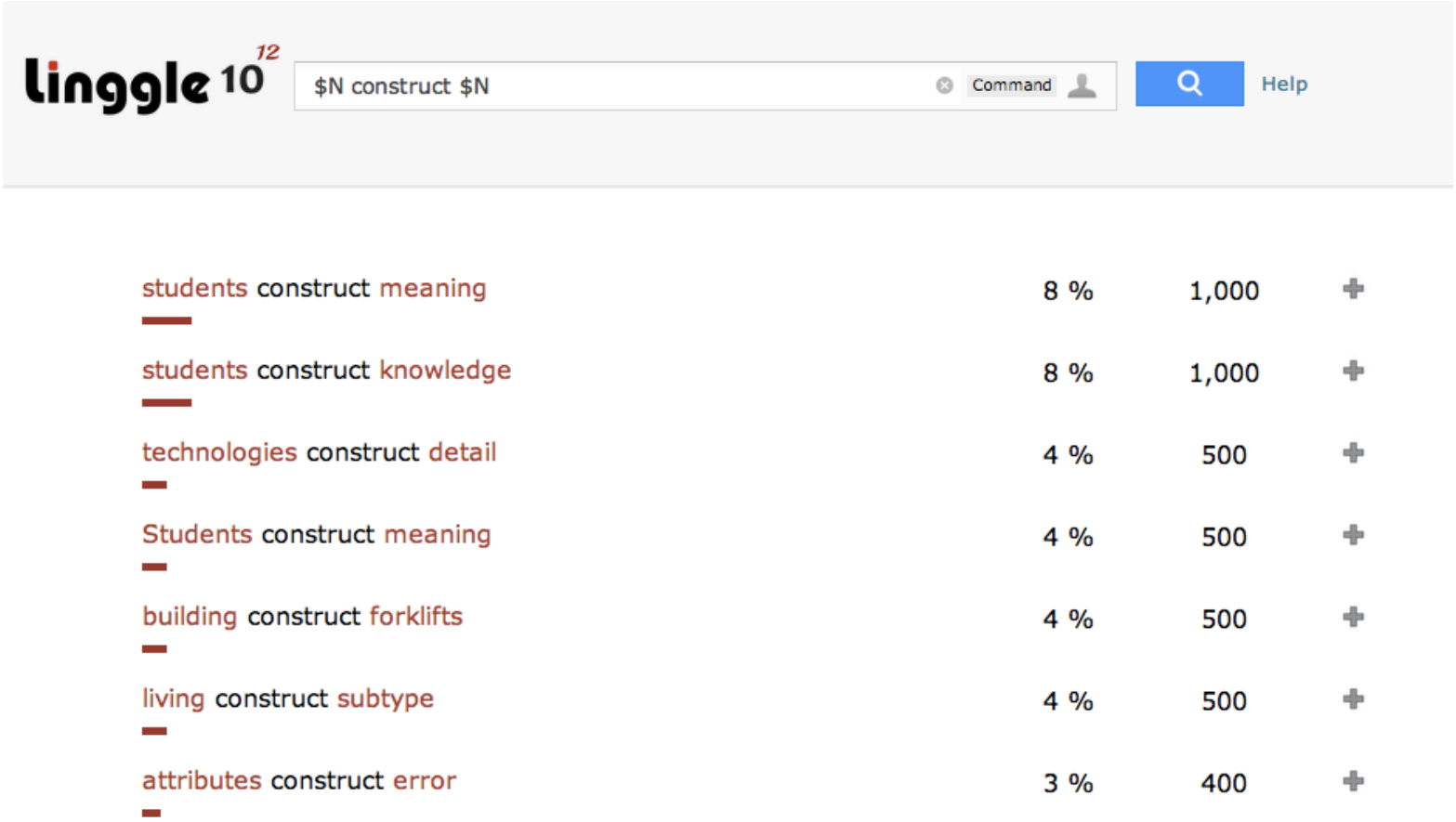
Human build Relationship

Human build Building | Machine

Hanks 2004

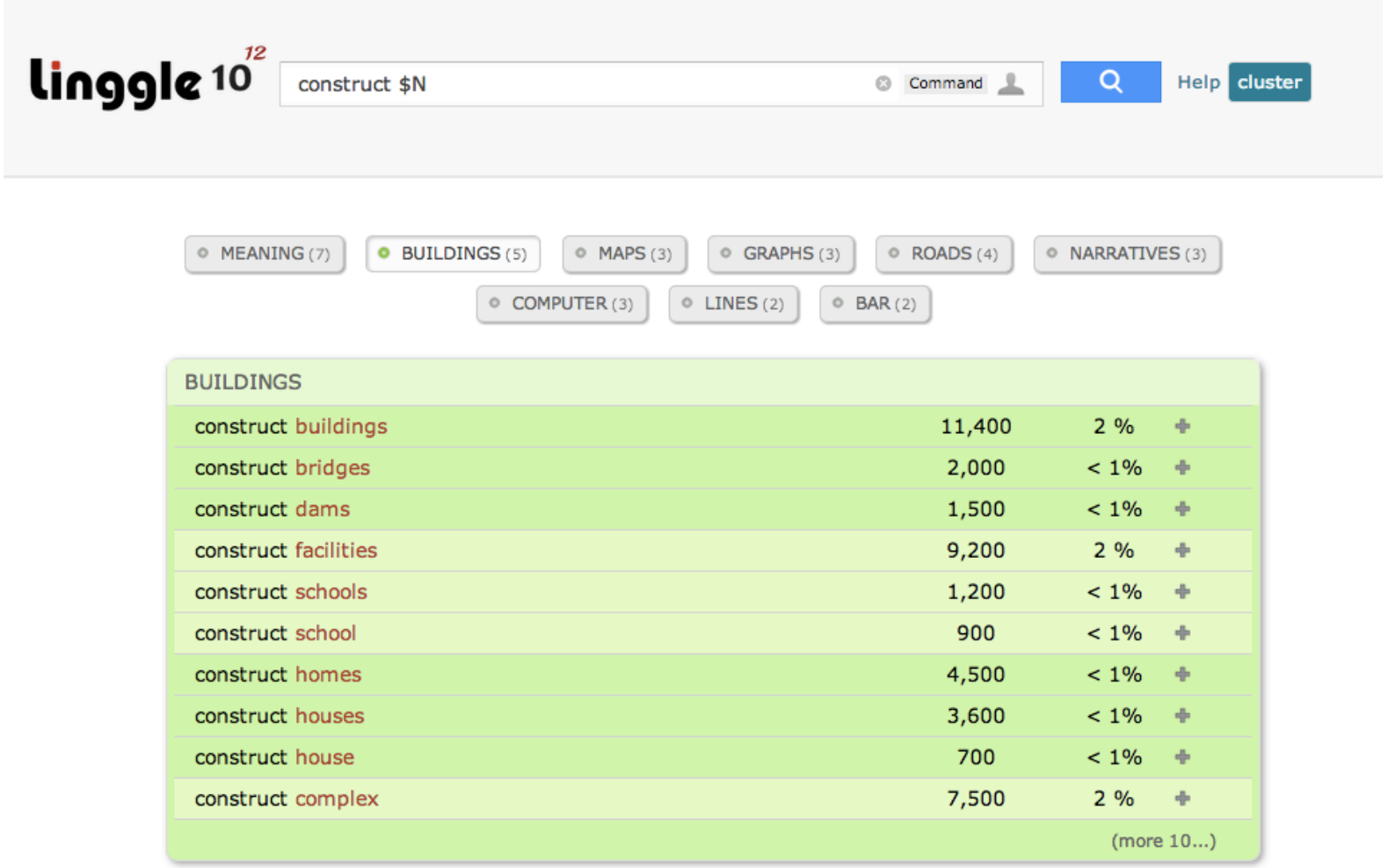
Requires expert linguists to manually derive patterns from a corpus

1. time-consuming
2. might not achieve high coverage for many verbs a learner has to master



Linggle: a Web-scale Linguistic Search Engine for Words in Context

- Boisson et al. 2013



Linggle website: linggle.com

Learners need more complete verb frames or patterns

Linggle →

1	**PERSON apologize**
	I apologize
	juvenile apologize
2	**PERSON apologize to PERSON**
	you apologize to me
	I apologize to him
3	**PERSON apologize for ACT**
	I apologize for needing
	I apologize for using

-FrameFinder

An automatically generating verb frames approach

1. Extract verb arguments - corpus
2. Obtain probable semantic roles for each argument - Knowledge base
3. Disambiguate senses of an argument - WSD
4. Tally and output verb frames

Step 1. Extracting verb arguments based on grammatical relations

"The deer would eat your plants"

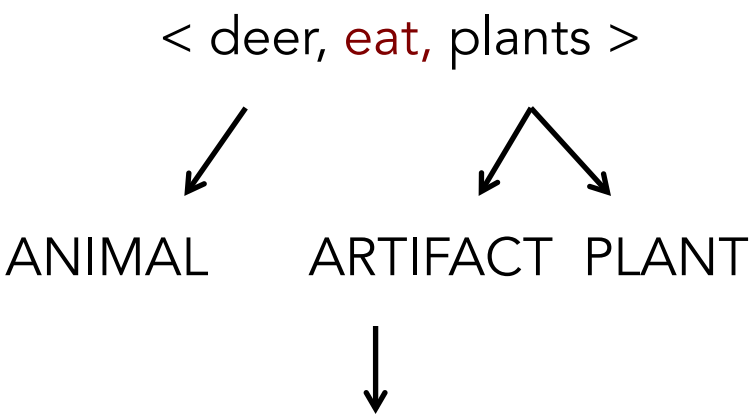
↓

(subject) "The deer would eat your plants" (object)

↓

< deer, eat, plants > (argument tuple)

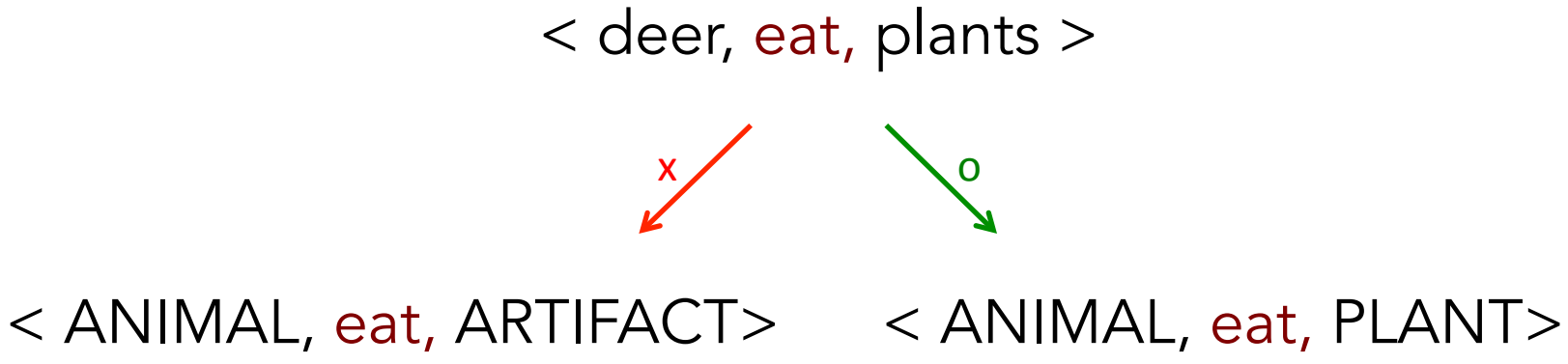
Step 2. Obtaining semantic roles from a knowledge base



Step 3. Generating verb frames

argument tuples	*semantic tuples*
deer, eat, plants	< ANIMAL, eat, PLANT >
	< ANIMAL, eat, ARTIFACT >
birds, eat, seeds	< ANIMAL, eat, PLANT >
	< ANIMAL, eat, PERSON>
ants, eat, sugar	< ANIMAL, eat, FOOD>

Disambiguate and identify the *intended meaning* of arguments



Applying the Expectation-Maximization (EM) algorithm to disambiguate

- EM algorithm - Dempster et al., 1977
- Expectation (E-step)
- Maximization (M-step)

A flexible and extensible disambiguation module

$$p(v_f|v) = \alpha(\, p(v_f|v)\,) + \beta(\, p(s|v) \cdot p(o|v) \cdot \prod_{po\ \epsilon\ adverbial} p(po|v)\,)$$

subject object adverbial

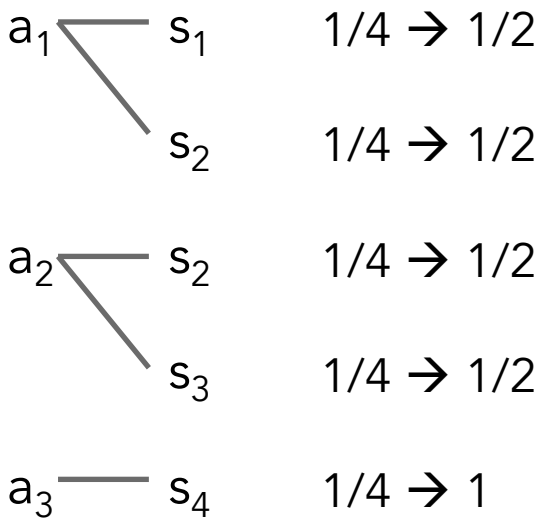
In this study, we set $\alpha = 1,\ \beta = 0$

```
Procedure GenerateVerbFrames(ArgumentTuples)
 (1a)  FrameCandis = {}
 (1b)  allFrames = {getVerbFrames(t) for each tuple t in ArgumentTuples}
 (1c)  p(frame|v) = 1 / |allFrames|
       while not convergent
         count(frame|v) = 0 for frame in allFrames
         for each tuple t in ArgumentTuples
 (2a)      tupleFrame = getVerbFrames(t)
 (2b)      nf = sum(prob(frame) for all frame in tupleFrames)
           for each frame in tupleFrames
 (2c)        count(frame) += prob(frame)/nf
         for all frame in allFrames
 (3)       p(frame|v) = count(frame|v)/|ArgumentTuples|
       for each tuple t in ArgumentTuples
 (4)     append argmax(frame) to FrameCandis
 (5)   RankedFrames = Sort(frames in FrameCandis in decreasing order of frequency)
     return top K ranking RankedFrames
```

EM Example

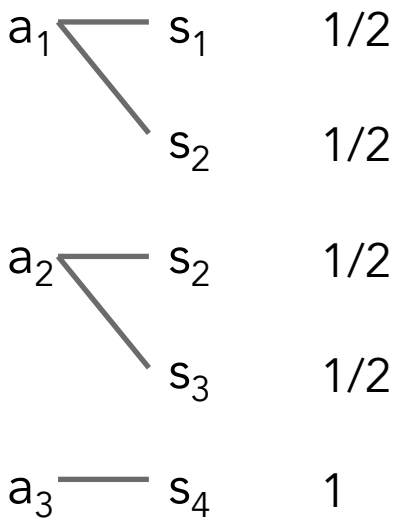
argument tuples	*semantic tuples*	
deer, eat, plants (a_1)	< ANIMAL, eat, ARTIFACT >	s_1
	< ANIMAL, eat, PLANT>	s_2
birds, eat, seeds (a_2)	< ANIMAL, eat, PLANT >	s_2
	< ANIMAL, eat, PERSON>	s_3
ants, eat, sugar (a_3)	< ANIMAL, eat, FOOD>	s_4

1st E-step: Initialize the probability of each candidate *uniformly*



→ normalize

1st M-step:



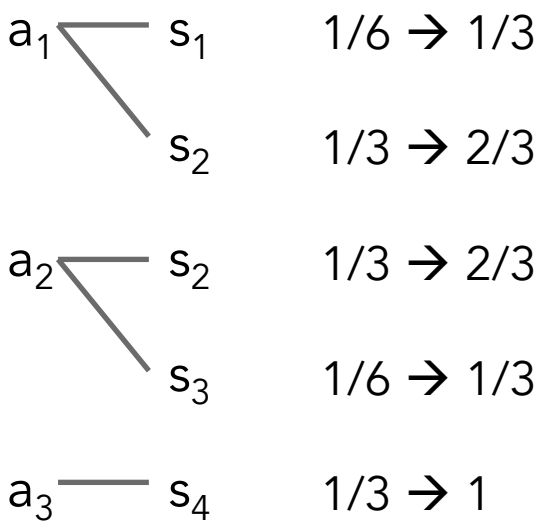
s_1 : 1/2 /3 = 1/6

s_2 : (1/2 + 1/2) /3 = 1/3

s_3 : 1/2 /3 = 1/6

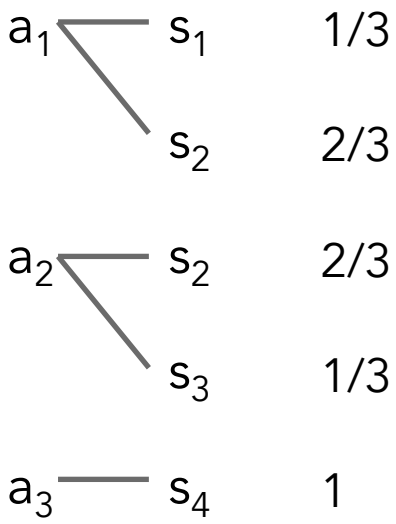
s_4 : 1 /3 = 1/3

2nd E-step:



→ normalize

2nd M-step:



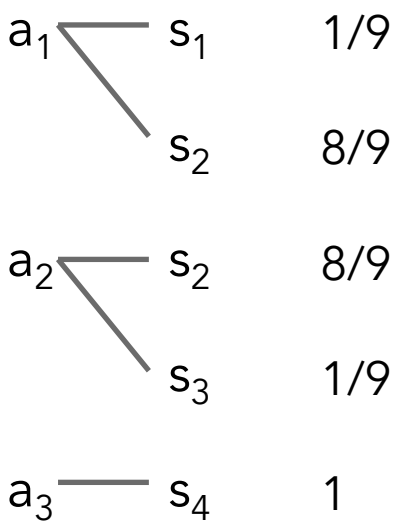
s_1 : 1/3 /3 = 1/9

s_2 : (2/3 + 2/3) /3 = 4/9

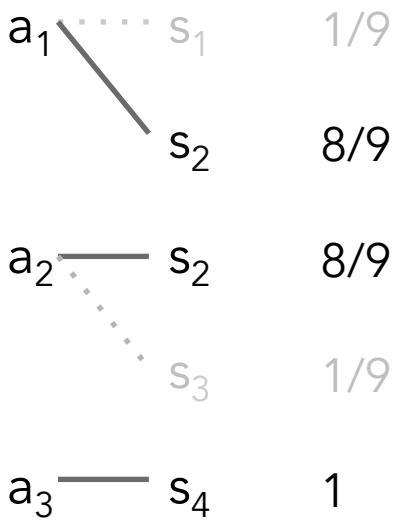
s_3 : 1/3 /3 = 1/9

s_4 : 1 /3 = 1/3

Until convergent…



Assign a semantic tuple for each argument tuple



Finally, counts and outputs the ranked semantic tuples

ANIMAL eat PLANT	s_2	2
ANIMAL eat FOOD	s_4	1
< ANIMAL, eat, ARTIFACT >	s_1	0
< ANIMAL, eat, PERSON>	s_3	0

Verb frame generation methods compared

- Most Frequent Sense (**MFS**)

  - always chooses the major sense

    e.g., < deer, eat, plants > → < ANIMAL, eat, ARTIFACT >

- Expectation-Maximization: (**EM**)

  - considers probabilities of semantic categories

The extracted frames for the average language learners should be:

1. Valid
2. Cover common usages of the verb

Evaluate *FrameFinder*'s performance using

1. Precision
2. Weighted recall
3. F-measure

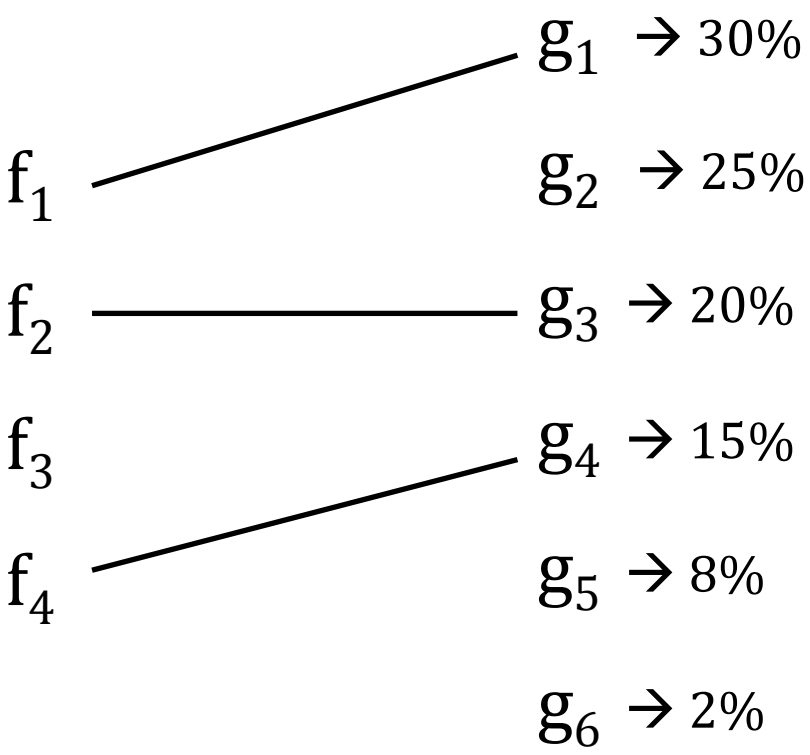
$$Precision(v) = \frac{|F(v) \cap Gold(v)|}{|F(v)|}$$

$$wtRecall(v) = \sum_{f \in F(v) \cap Gold(v)} \frac{freq(f)}{|Gold(v)|}$$

$$F\_measure = \frac{2PR}{P + R}$$

where $P = Precision$, $R = wtRecall$

For example, given a verb *v*



Precision = 0.6

wtRecall = 0.3+0.2+0.15

= 0.65

F-measure = 0.62

Resources and tools we used

- British National Corpus (BNC)
  - 4,693,767 sentences
- Stanford Parser
  - subject, object, adverbial- p.22 Table 5
- WordNet
  - 25 supersenses - p.22 Table 4
- Corpus Pattern Analysis (CPA)
  - 812 verbs
  - 3,100 verb patterns

Mapping CPA labels to WordNet supersenses

e.g.,

HUMAN | INSTITUTION abandon PLAN | ACTIVITY

PERSON abandon COGNTION

PERSON abandon ACT

GROUP abandon COGNITION

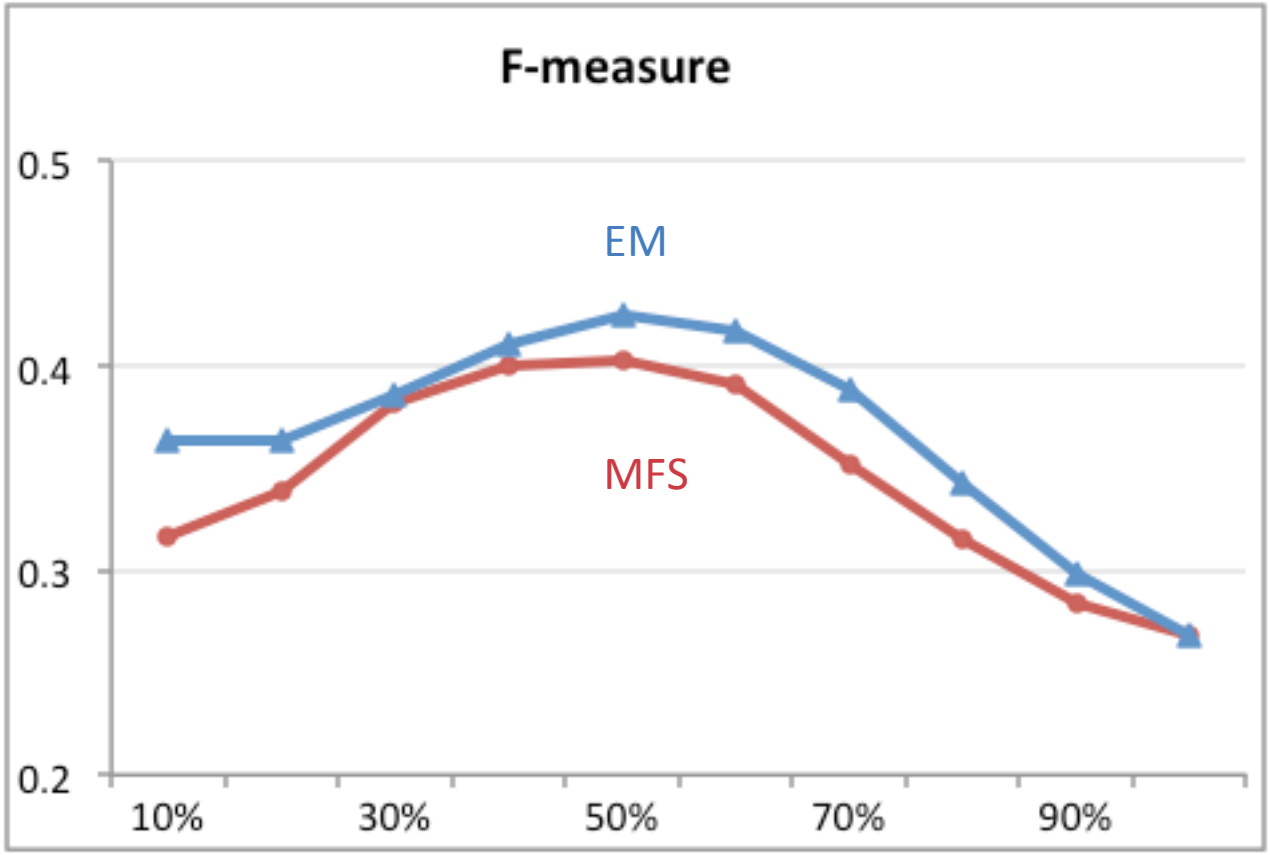
GROUP abandon ACT

Calculate the coverage

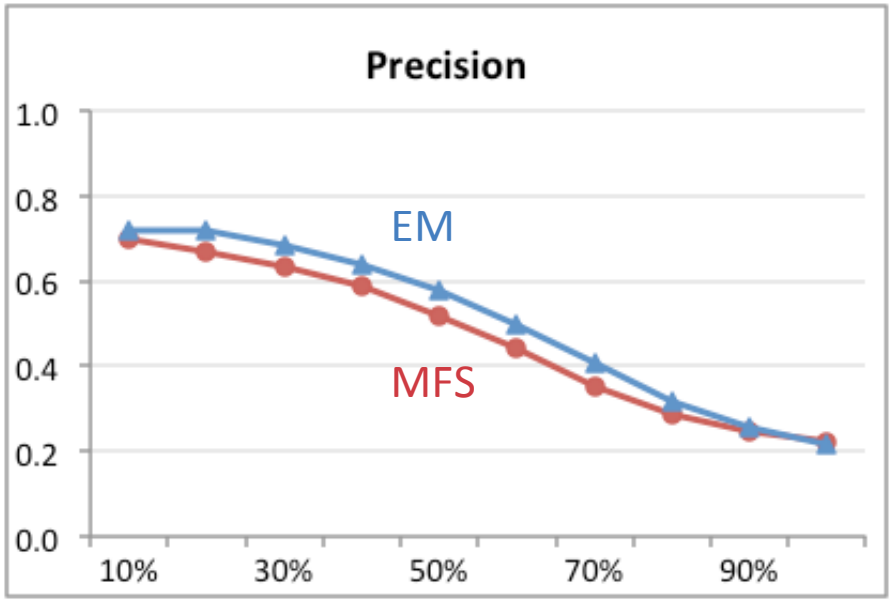
Verb frame	Coverage	Accumulated coverage
PERSON abandon COGNITION	28.3%	28.3%
PERSON abandon ACT	20.2%	48.5%
PERSON abandon PERSON	18.18%	66.7%
⋮	⋮	⋮

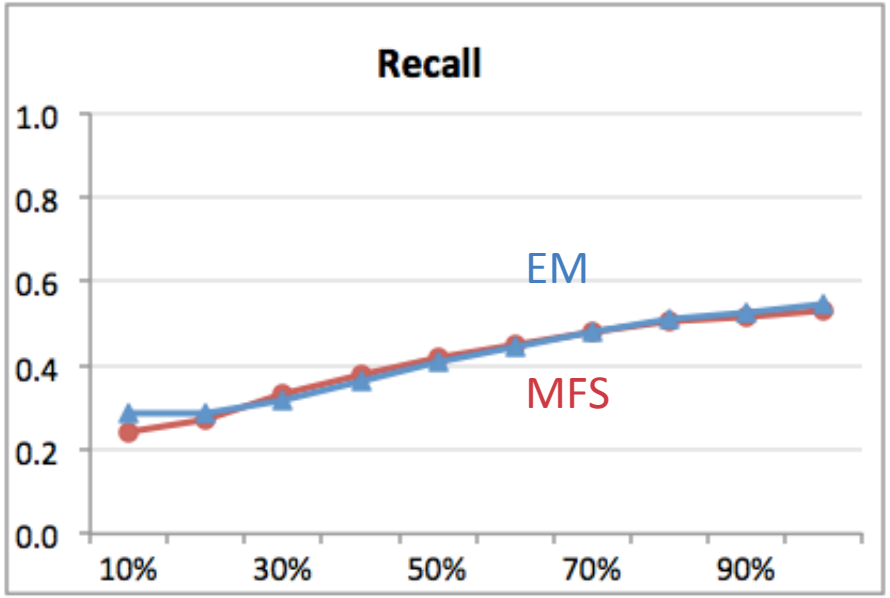
40%

Overall performance of *FrameFinder*

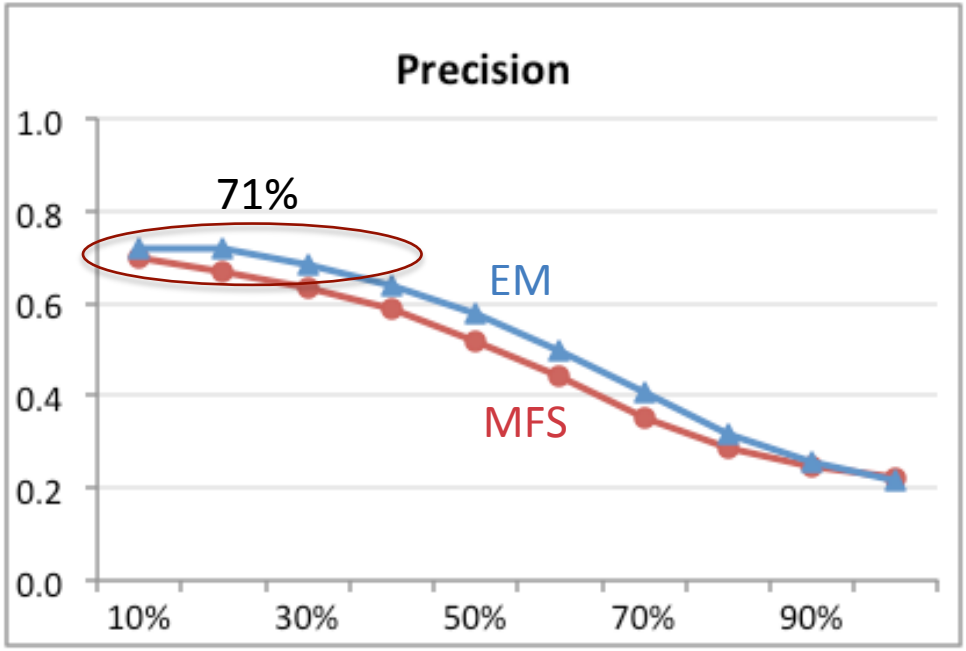


Precision and Recall of *FrameFinder*



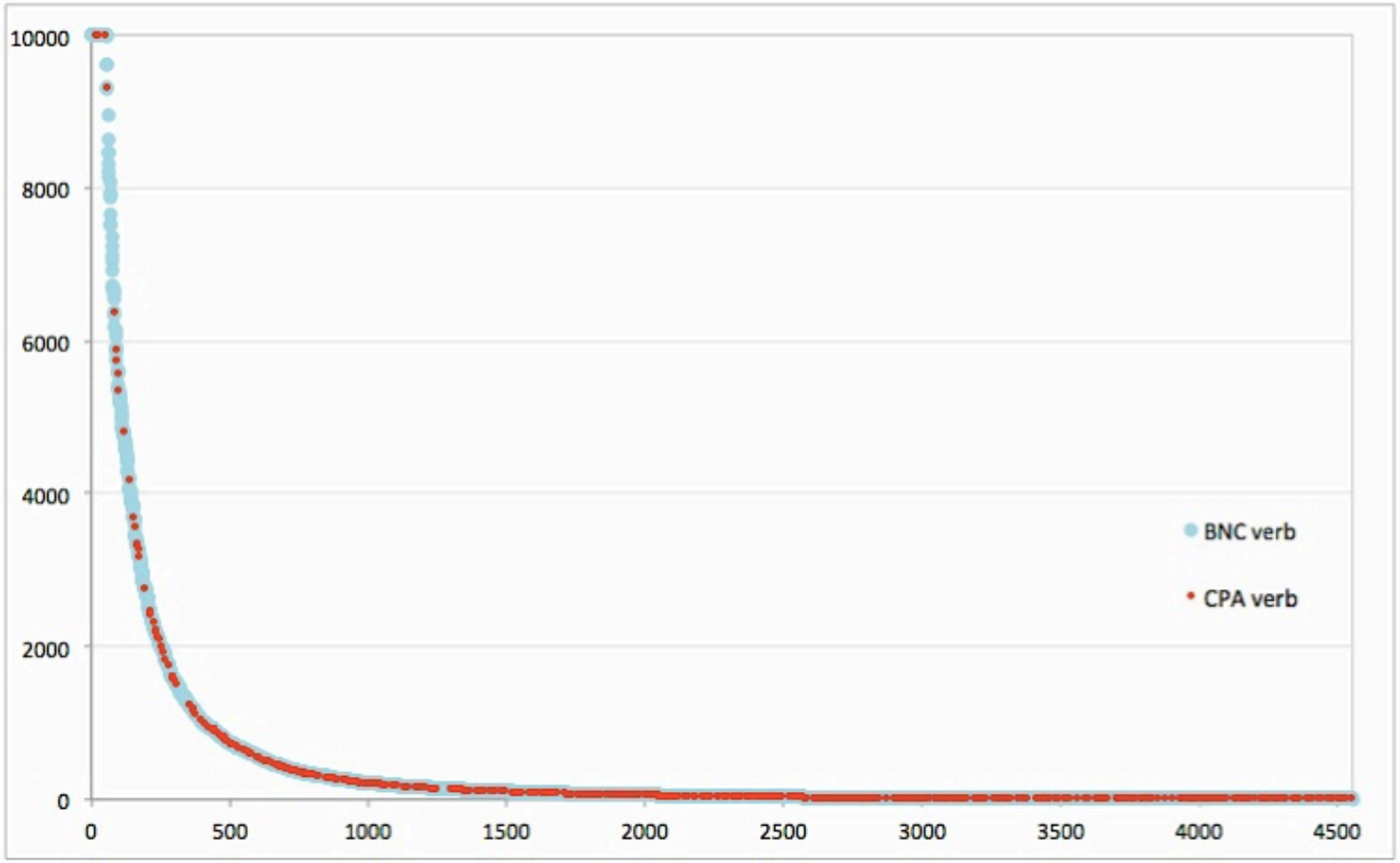


Our goal is to provide verb frames to facilitate language learning



We need high precision rate in high frequency verbs

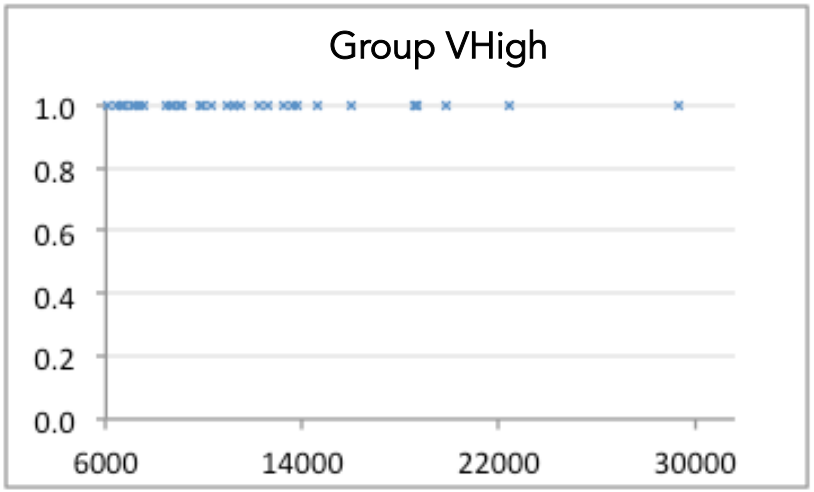
Number of occurrences of verbs in BNC and CPA

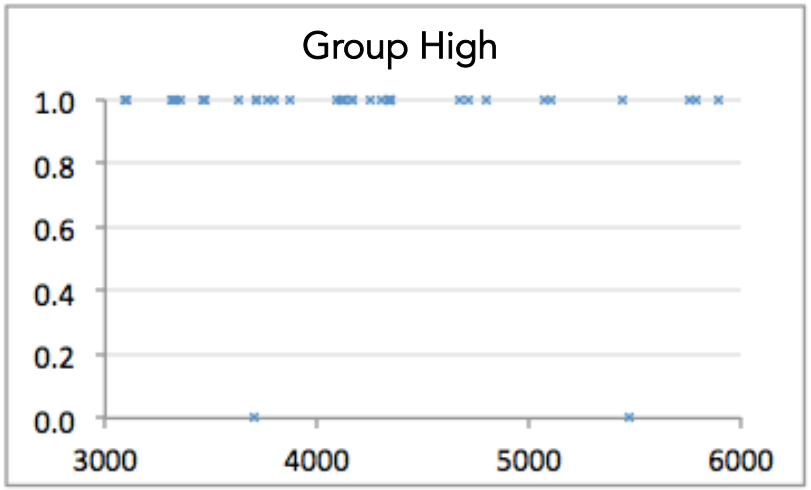


Five grouping criteria and the number of verbs in CPA

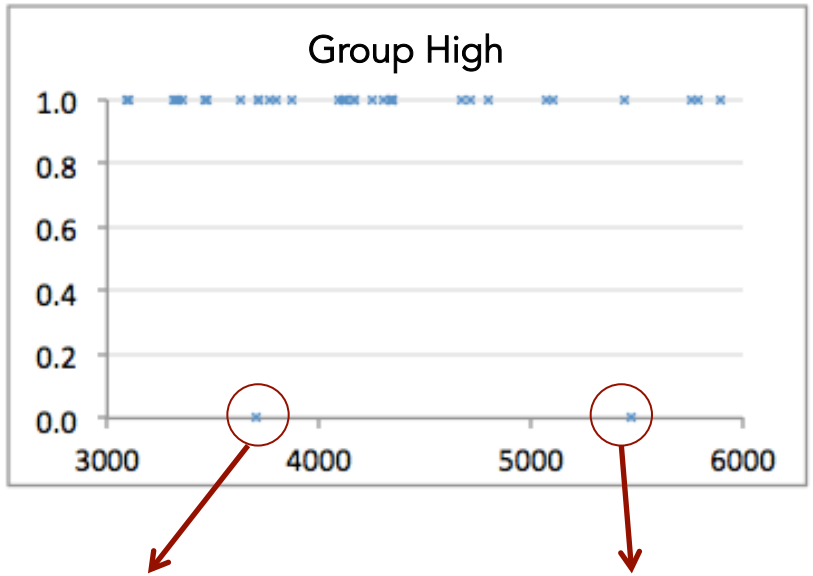
Group	Verb count criterion	# of verbs in CPA	Verb samples
VHigh	> 6000	26 → 3%	talk, say, tell
High	3000 - 6000	50 → 6%	propose, accuse, drink
Mid	500 - 3000	85 → 10%	abolish, dispose, pray
Low	150 - 500	77 → 9%	irritate, disregard, overflow
VLow	< 150	575 → 72%	petrify, abase, apostrophize

The precision of each verb in **VHigh** and **High**





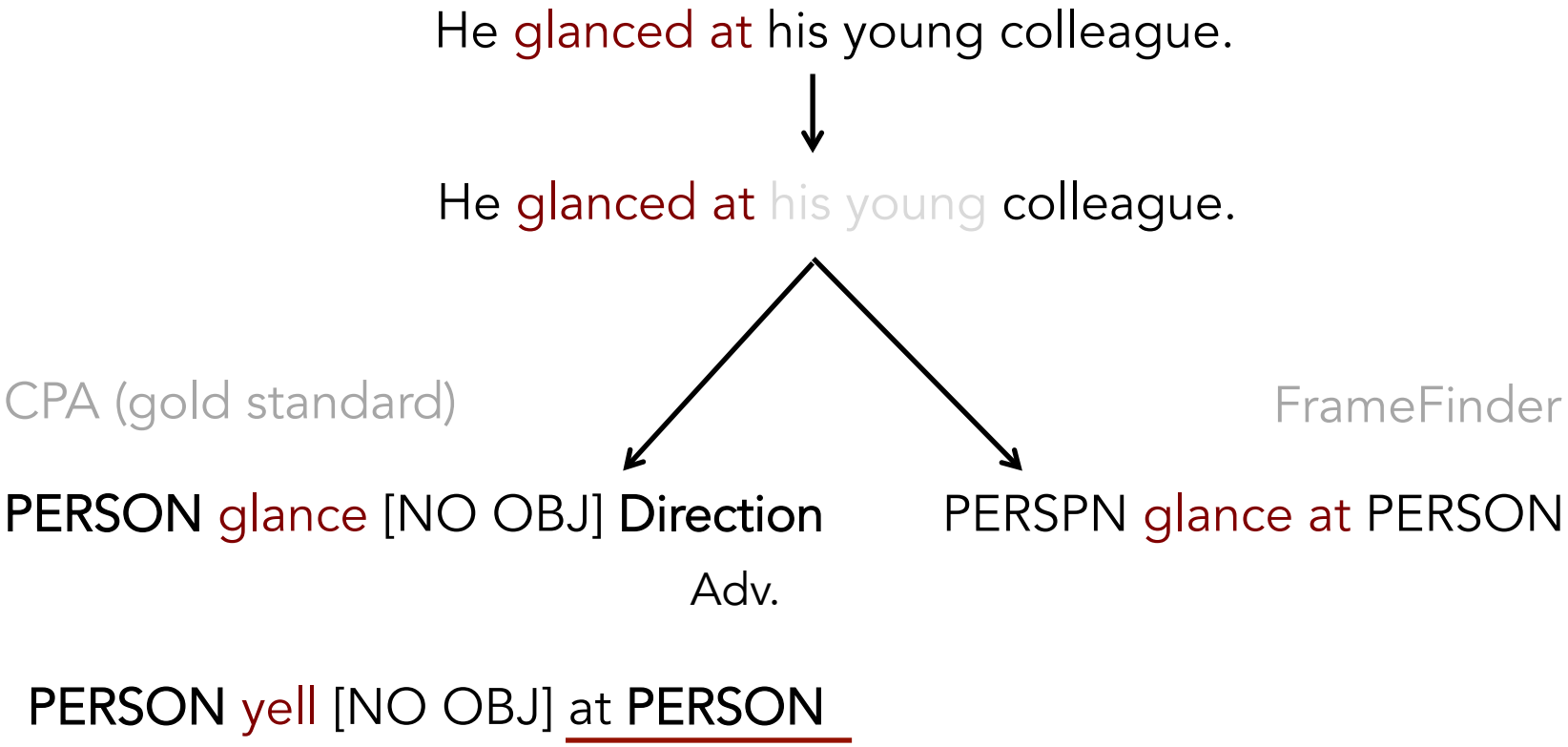
Failure cases in **Group High**



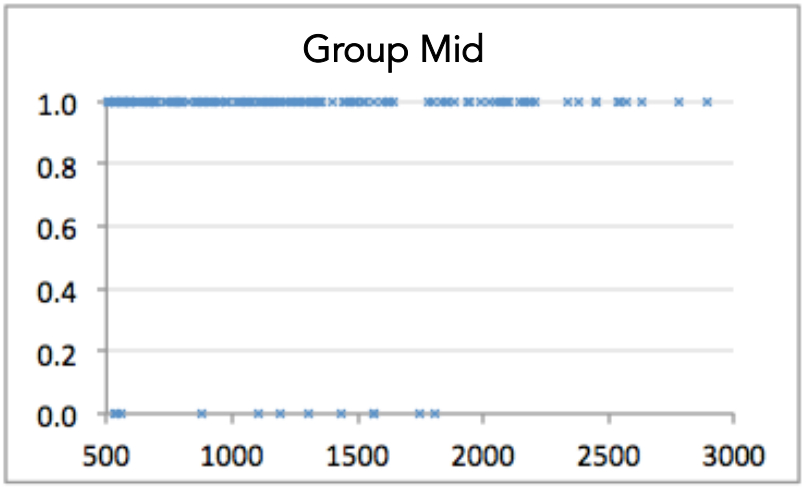
PERSON recall

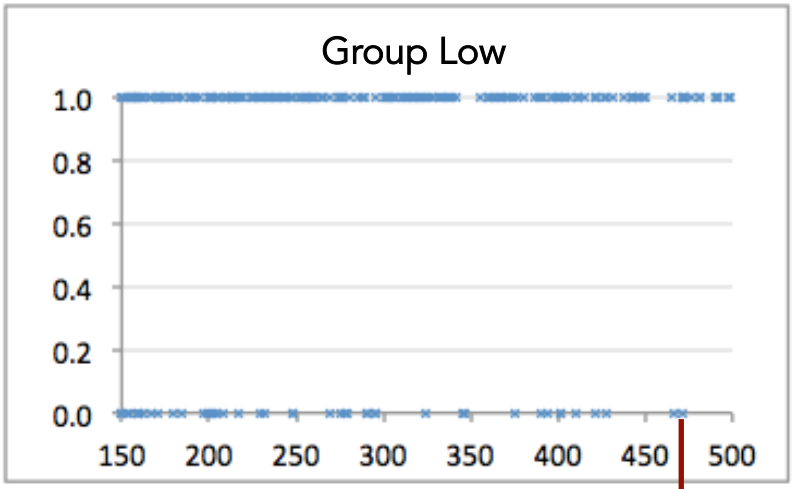
PERSON glance at **PERSON**

CPA internal labeling inconsistency



The precision of each verb in **Mid** and **Low**





anger

CPA-WordNet mapping inconsistency

The announcement is bound to anger the fundamentalists.

The announcement is bound to anger the fundamentalists.

FrameFinder: COMMUNICATION anger PERSON

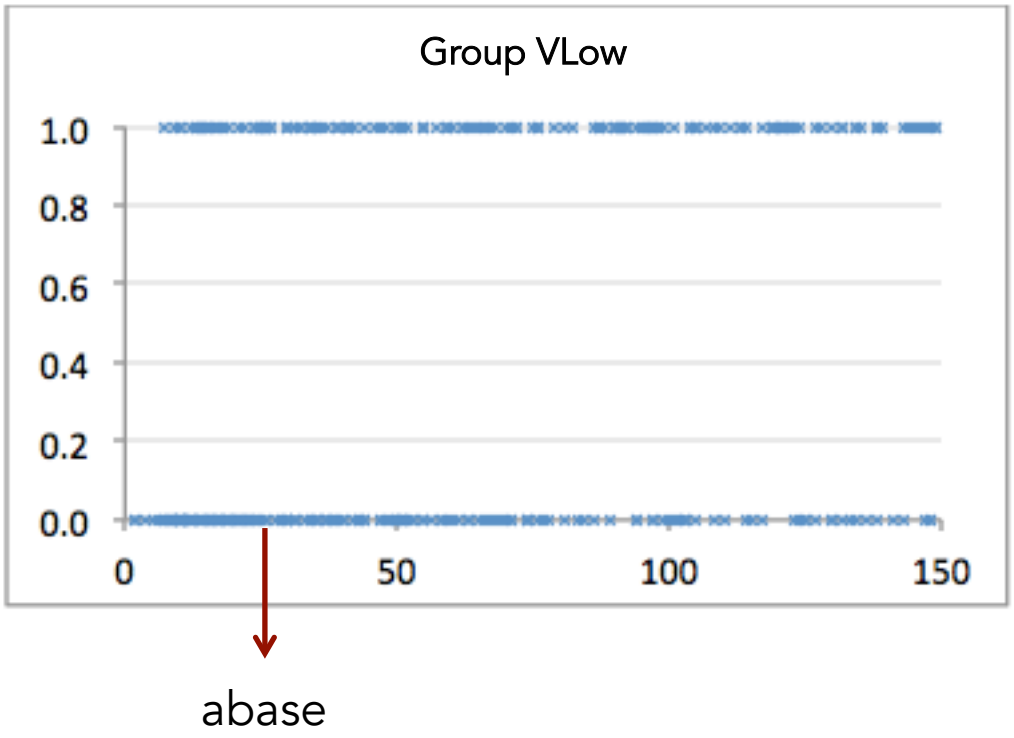
CPA-WordNet mapping inconsistency

FrameFinder: COMMUNICATION anger PERSON

CPA: [**Event**] anger PERSON

EVENT STATE

The precision of each verb in **VLow**



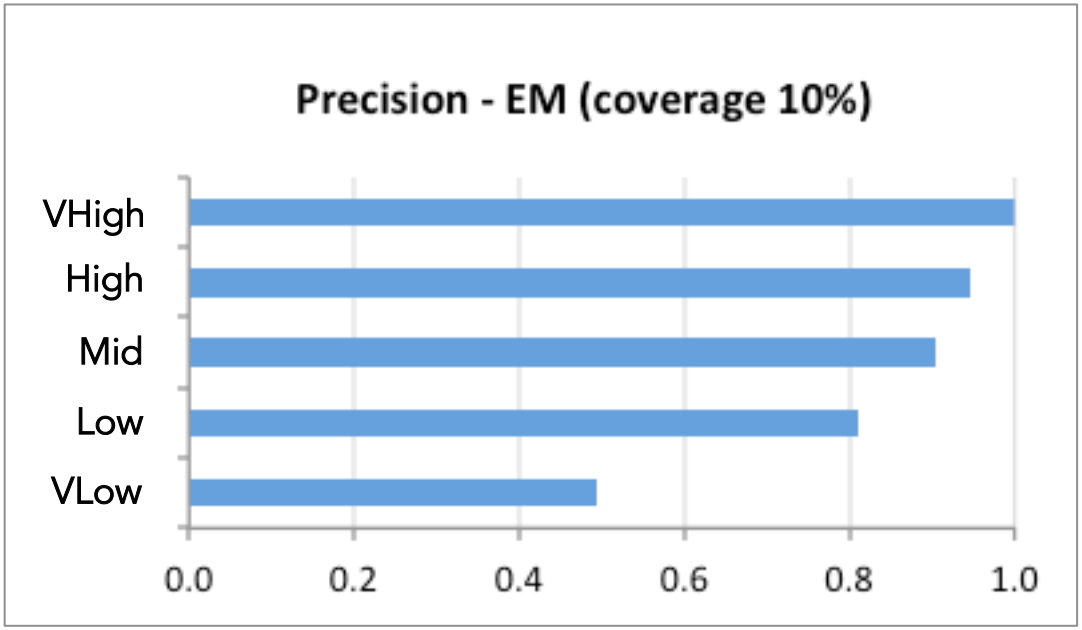
Optional adverbial

FrameFider: PERSON abase PERSON in STATE

PERSON abase PERSON before PERSON

CPA: PERSON abase PERSON

The precision of *FrameFinder* in each groups



Summary

- We have introduced a method combining corpus statistics and knowledge base to disambiguate and automatically acquire verb frames

- Evaluation results show that the method is able to extract verb frames with reasonable precision and recall of the most frequent usages for language learning

Future work

- Existing very large corpora can be exploited
  - UKWaC: 2 billion words
  - ClueWeb: 70 billion words
- More grammatical relations should be utilized
  - Passive voice
  - Clause
- The generated verb frames could be applied in other fields
  - Verb clustering
  - Grammatical error detection/correction
  - Learning or teaching the difference of verb usage

Thank you!