

Breaking the Resource Bottleneck for Multilingual Processing

*University of Edinburgh
IGK Summer School
September 6, 2004*

Rebecca Hwa
University of Pittsburgh
hwa@cs.pitt.edu

Supervised Learning

- Training examples are pairs of **problems** and **answers**
- For part-of-speech (POS) tagging: **word in context**, **POS tag** pairs
- For parsing: **sentence**, **parse tree** pairs
- For text categorization: **article**, **category** pairs

Manual Annotation Issues

- Guideline development
 - Laborious, time-consuming
 - Costly
 - Inconsistencies
 - Inter-annotator (dis)agreements
- Few widely-accepted annotated resources
- Data are typically **news stories** in **English**

Multilingual Processing

Language	Treebank	Total Dev. Time	Corpus Size	Parser Performance
English	Penn Treebank	5 years	1M words 40k sentences	90%
Chinese	Chinese Treebank v2	2 years	100K words 4k sentences	75%
	Chinese Treebank v4	4 years	400K words 15k sentences	~80%
Others (e.g., Hindi, Farsi)	?		?	?

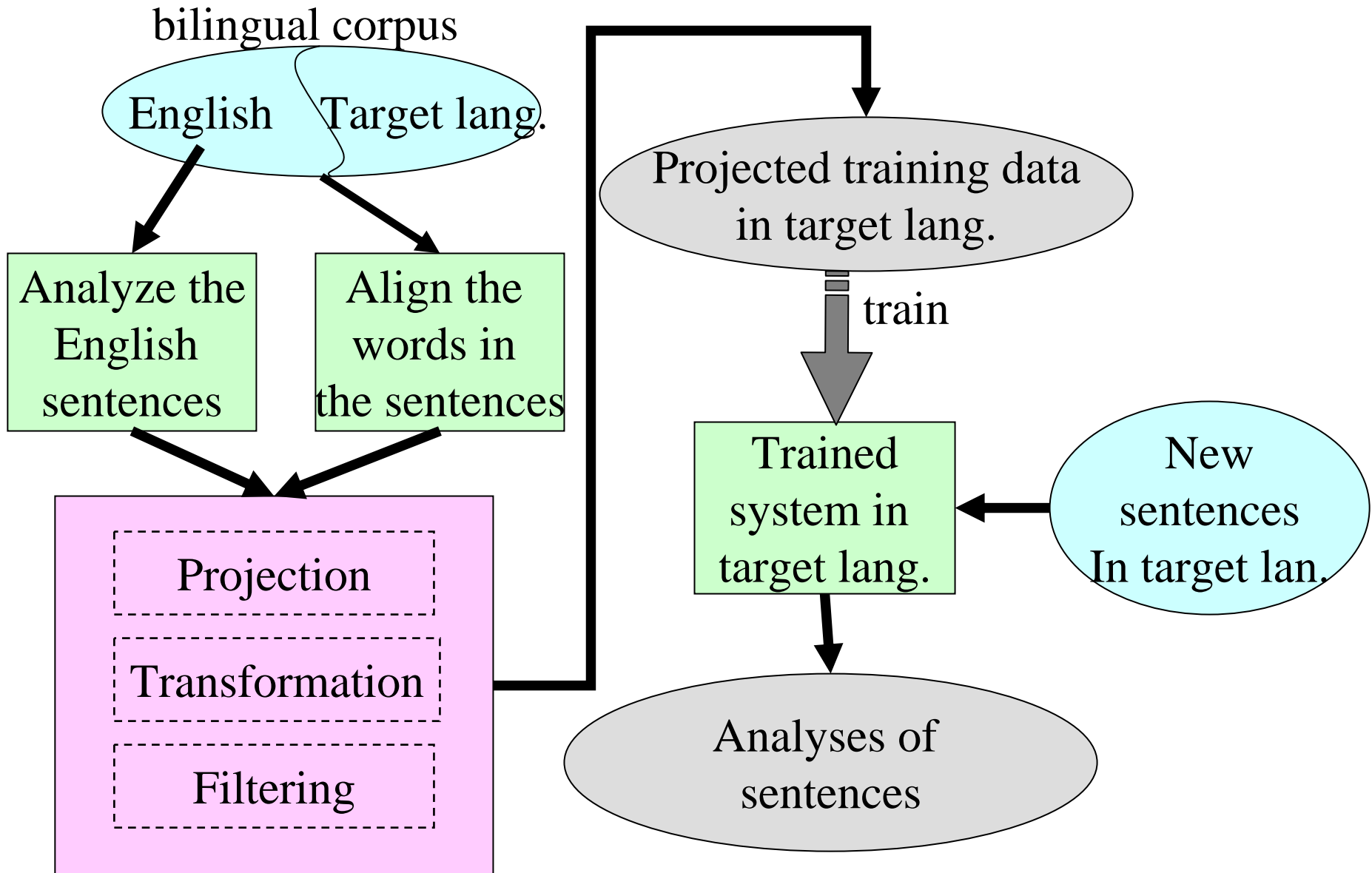
Research Questions

- How can annotated resources for non-English languages be acquired efficiently?
 - How to make use of resources we already have?
- How good are the resulting resources at training non-English systems?
 - Are the trained systems directly or indirectly useful?

Roadmap

- Motivation
- Annotation Projection
 - Overview
 - Theoretical Challenges
 - Practical Challenges
- Empirical studies
- Future Work

Projection Framework



Possible Applications of Projected Resources

- Morphological analyzer
- Base noun phrase chunker
- Name entity tagger
- POS tagger
- Syntactic dependency parser
- Semantic parser
- ...

Step 0. Acquire sentence aligned parallel text

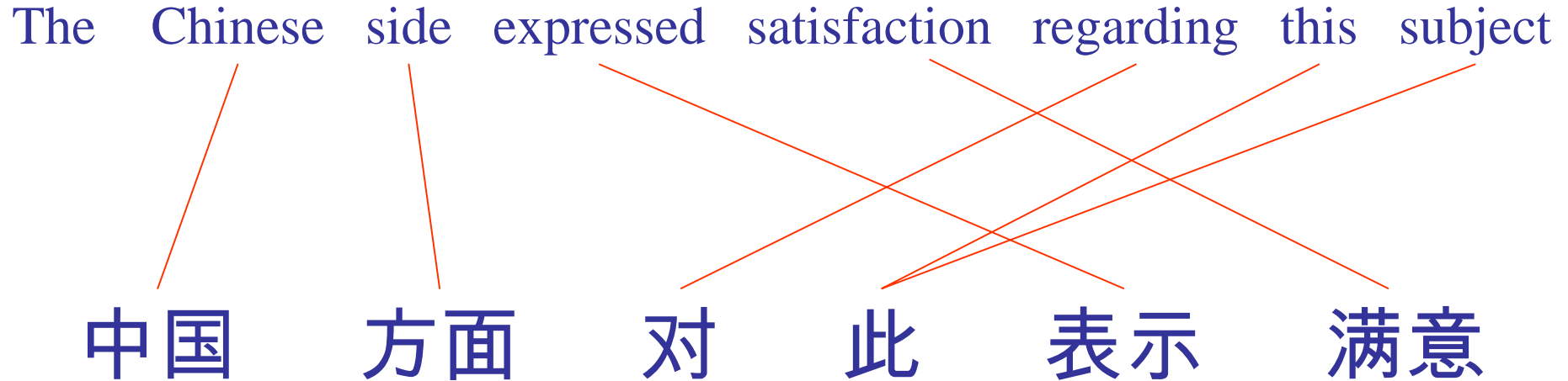
The Chinese side expressed satisfaction regarding this subject

中国 方面 对 此 表示 满意

Step 1. Align the Words

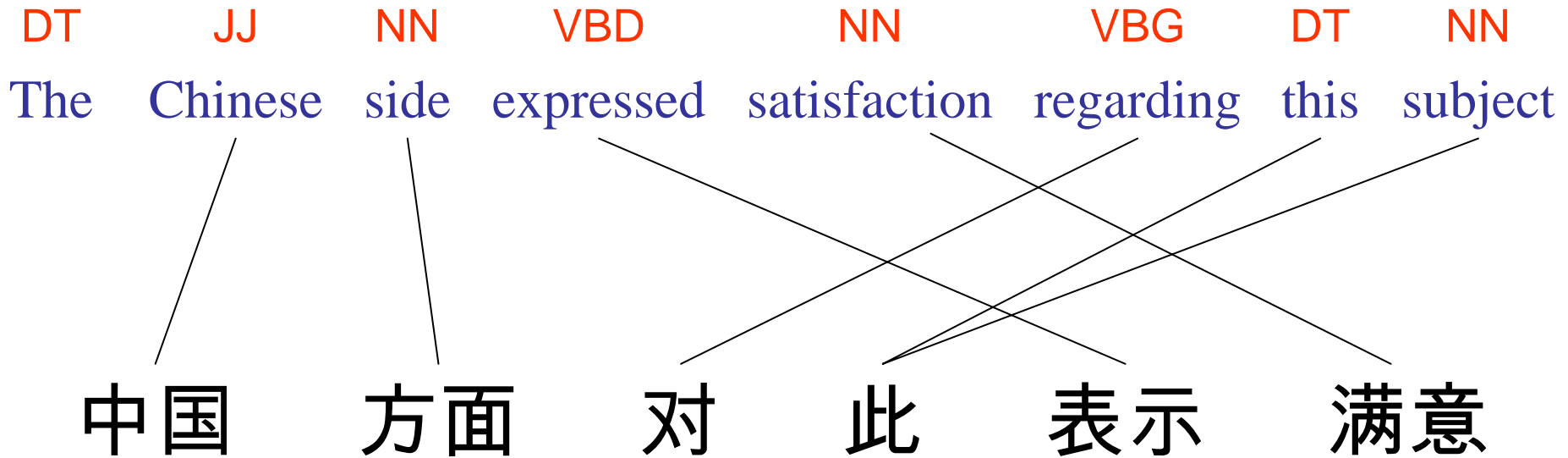
The Chinese side expressed satisfaction regarding this subject

中国 方面 对 此 表示 满意



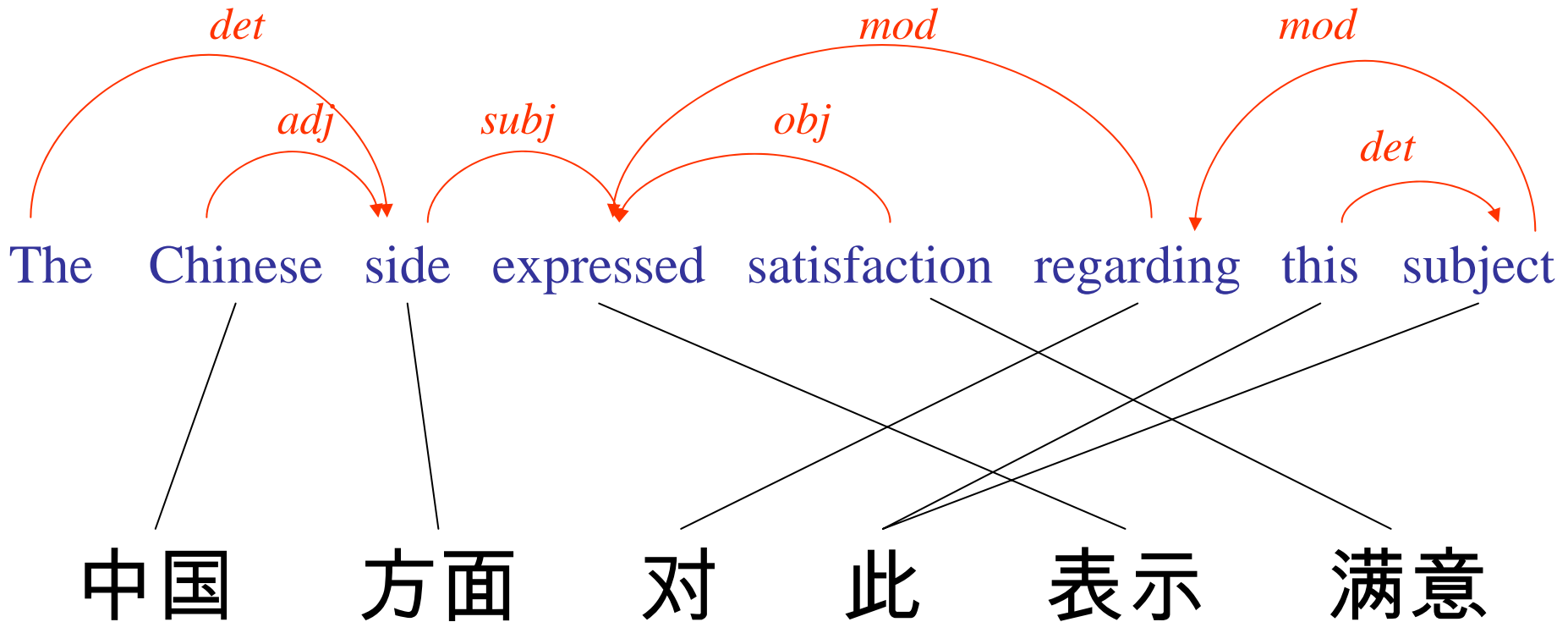
Step 2. Analyze the English Data

Example: POS Tagging



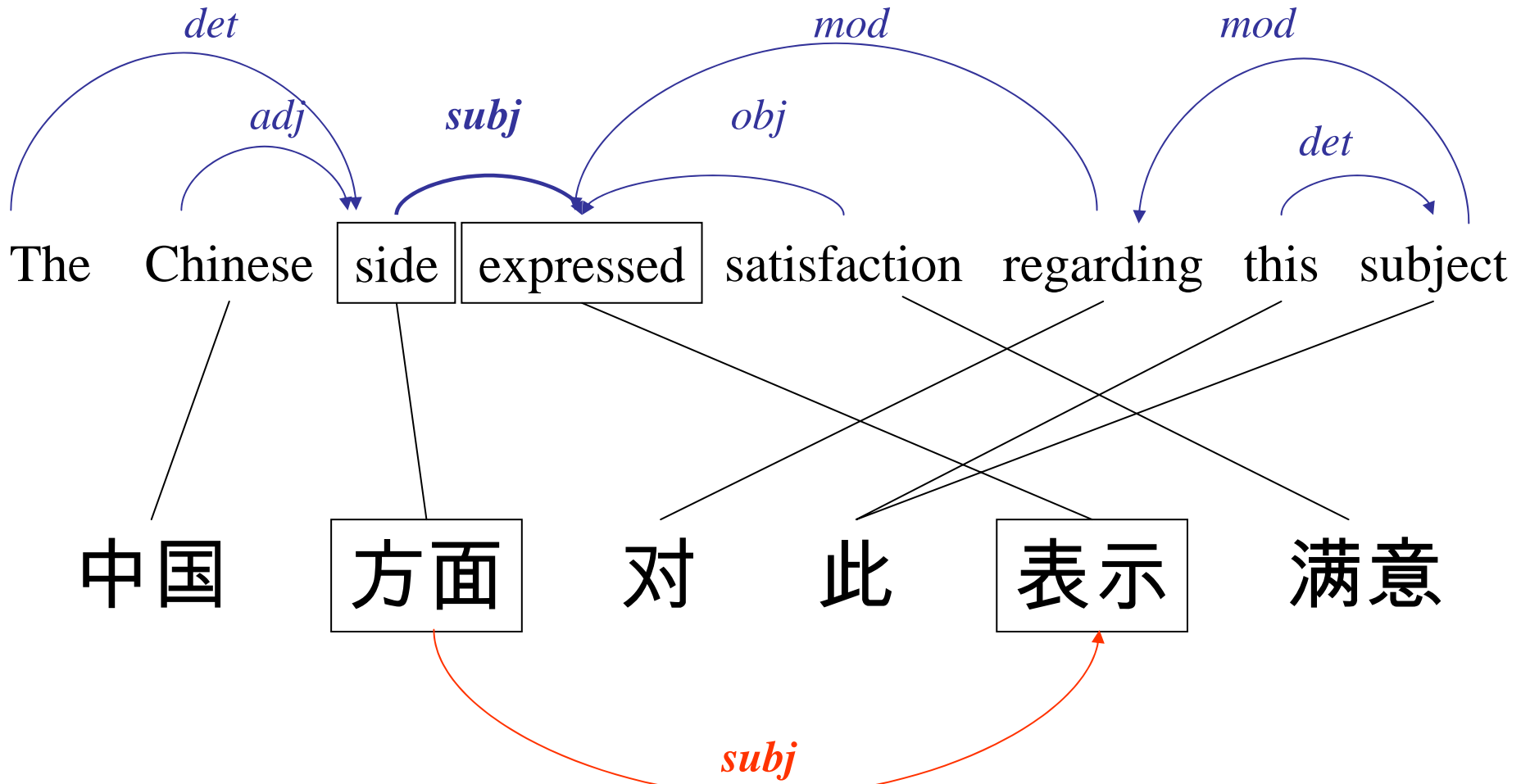
Step 2. Analyze the English Data

Example: Parsing



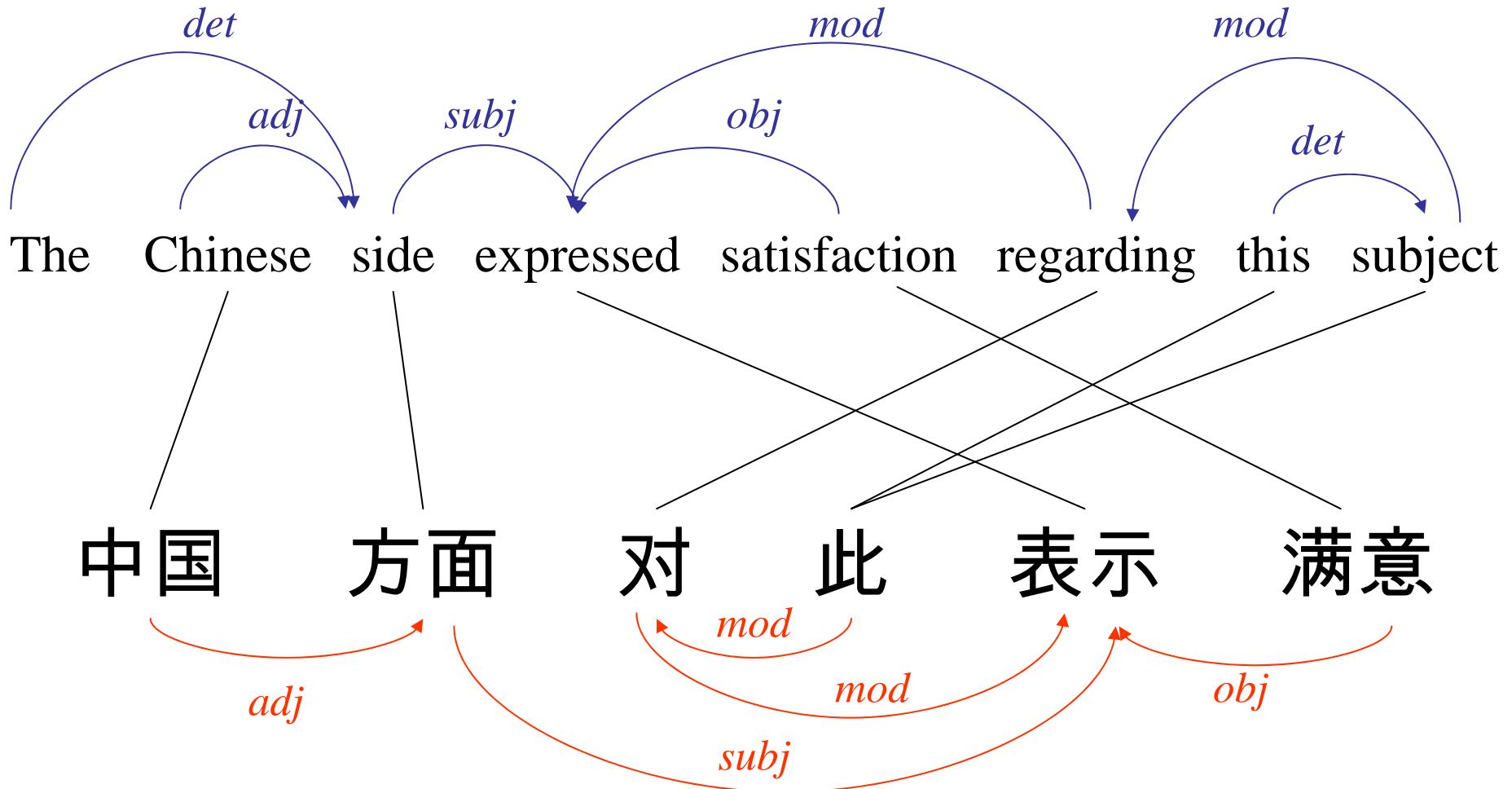
Step 3: Project Annotations across Alignments

Example: Parsing



Step 3: Project Annotations across Alignments

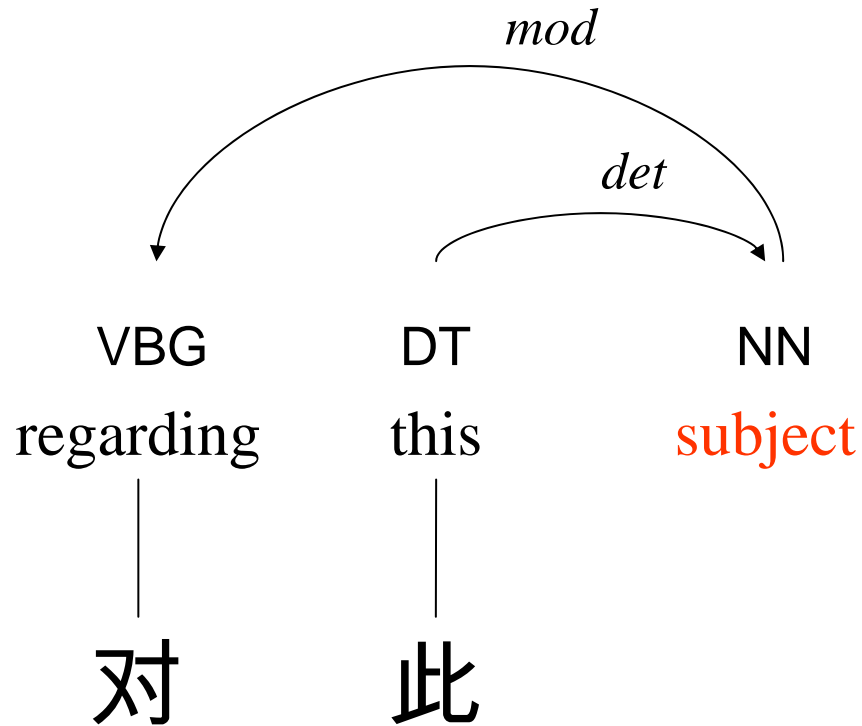
Example: Parsing



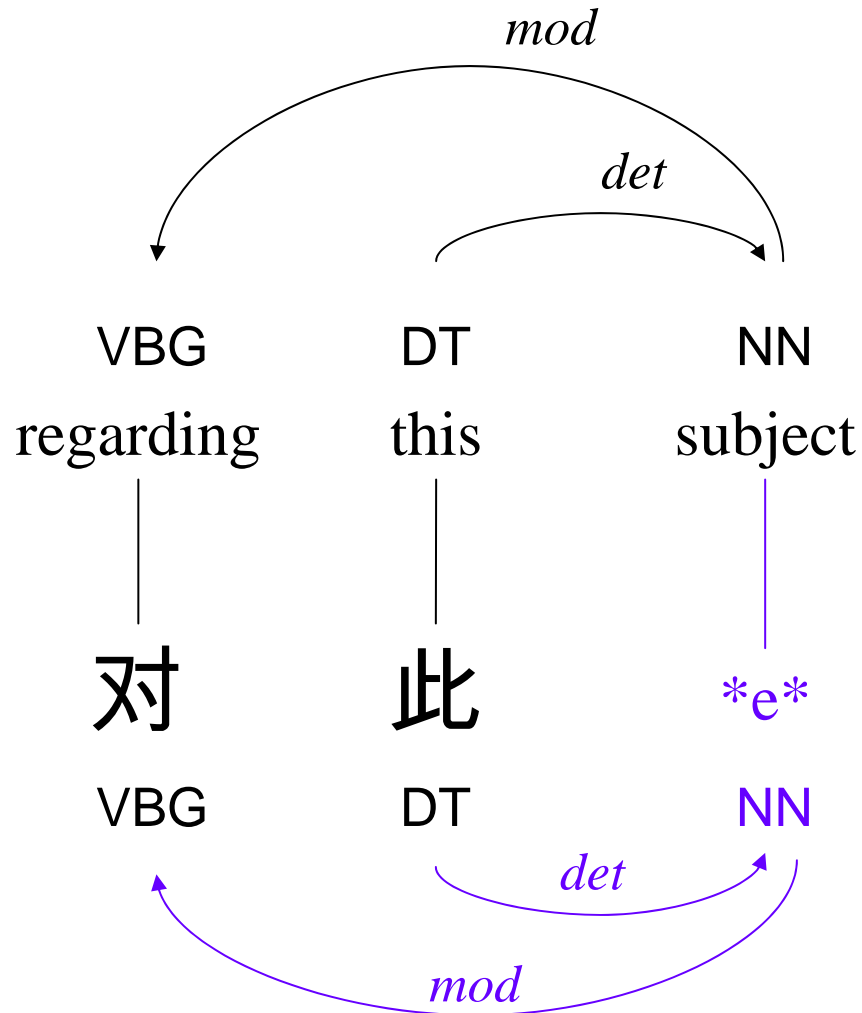
Challenges

- Theoretical challenge
 - Divergences: different languages express the same thing differently
 - Word alignments are not one-to-one
- Practical challenge
 - Framework relies on several components
 - Individual component errors can propagate

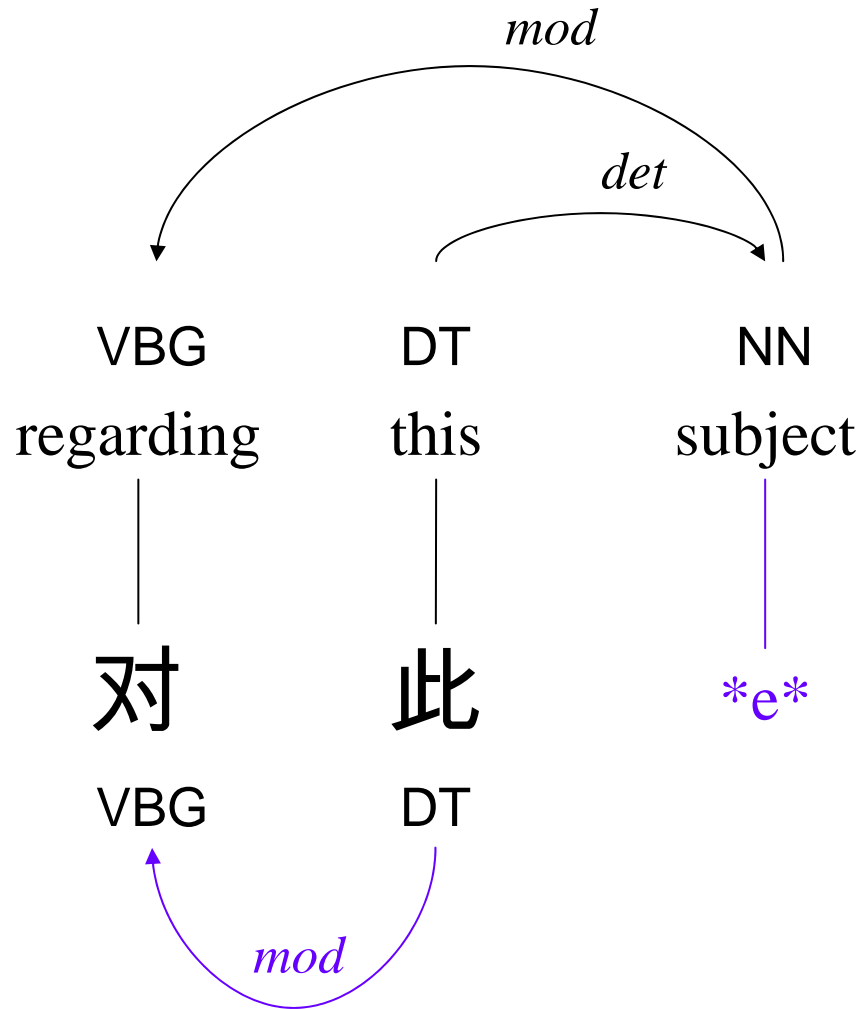
Unaligned English



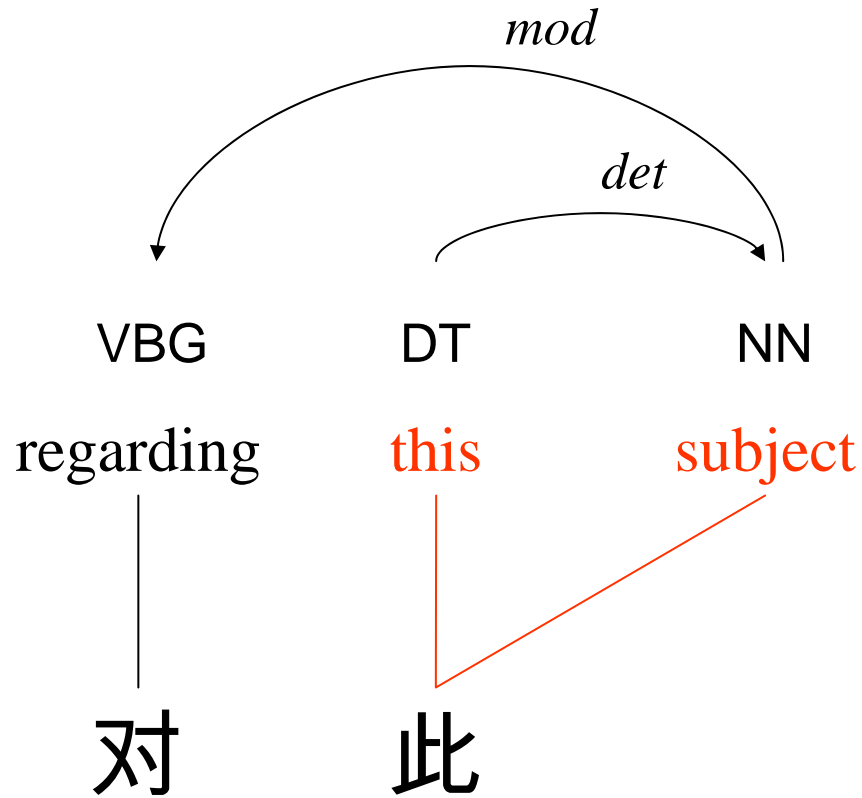
Unaligned English



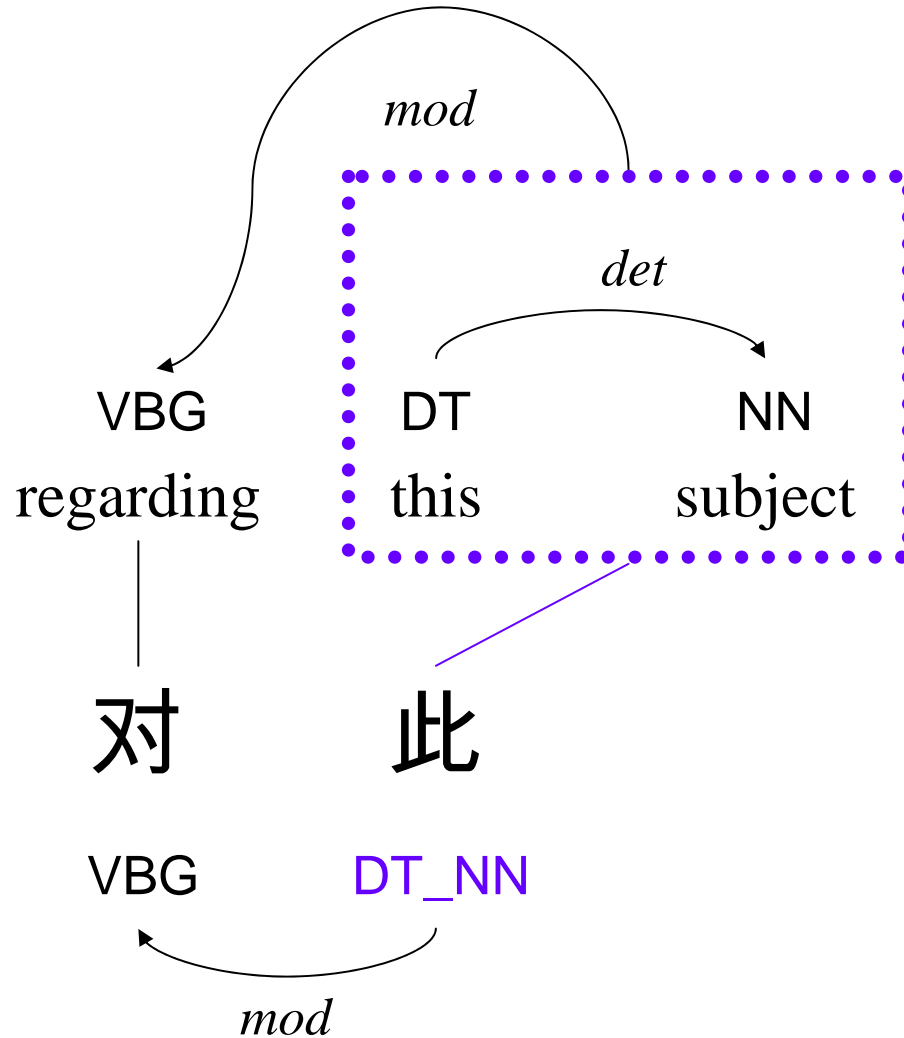
Unaligned English



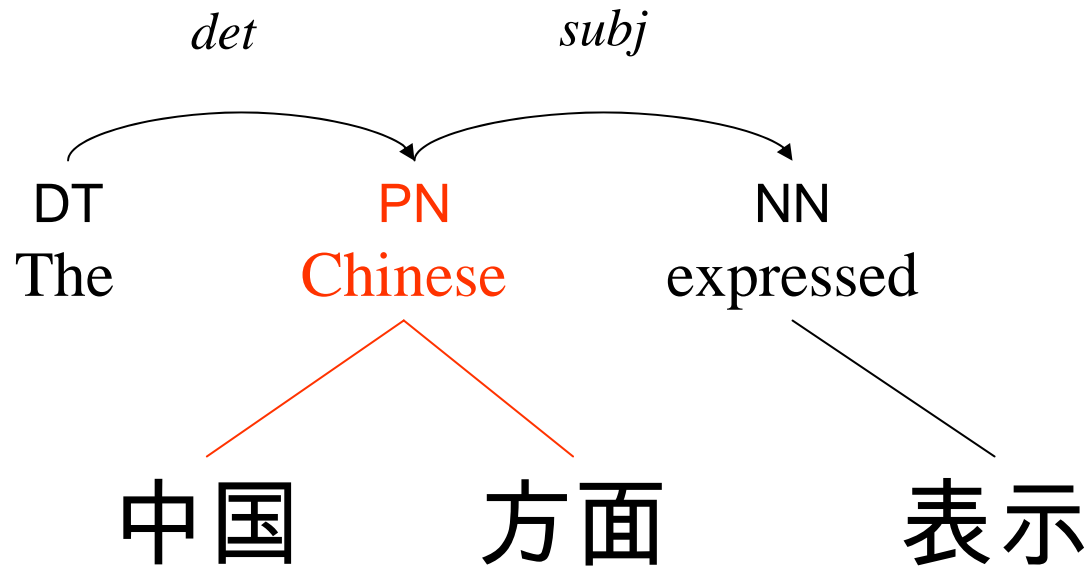
Many-to-1



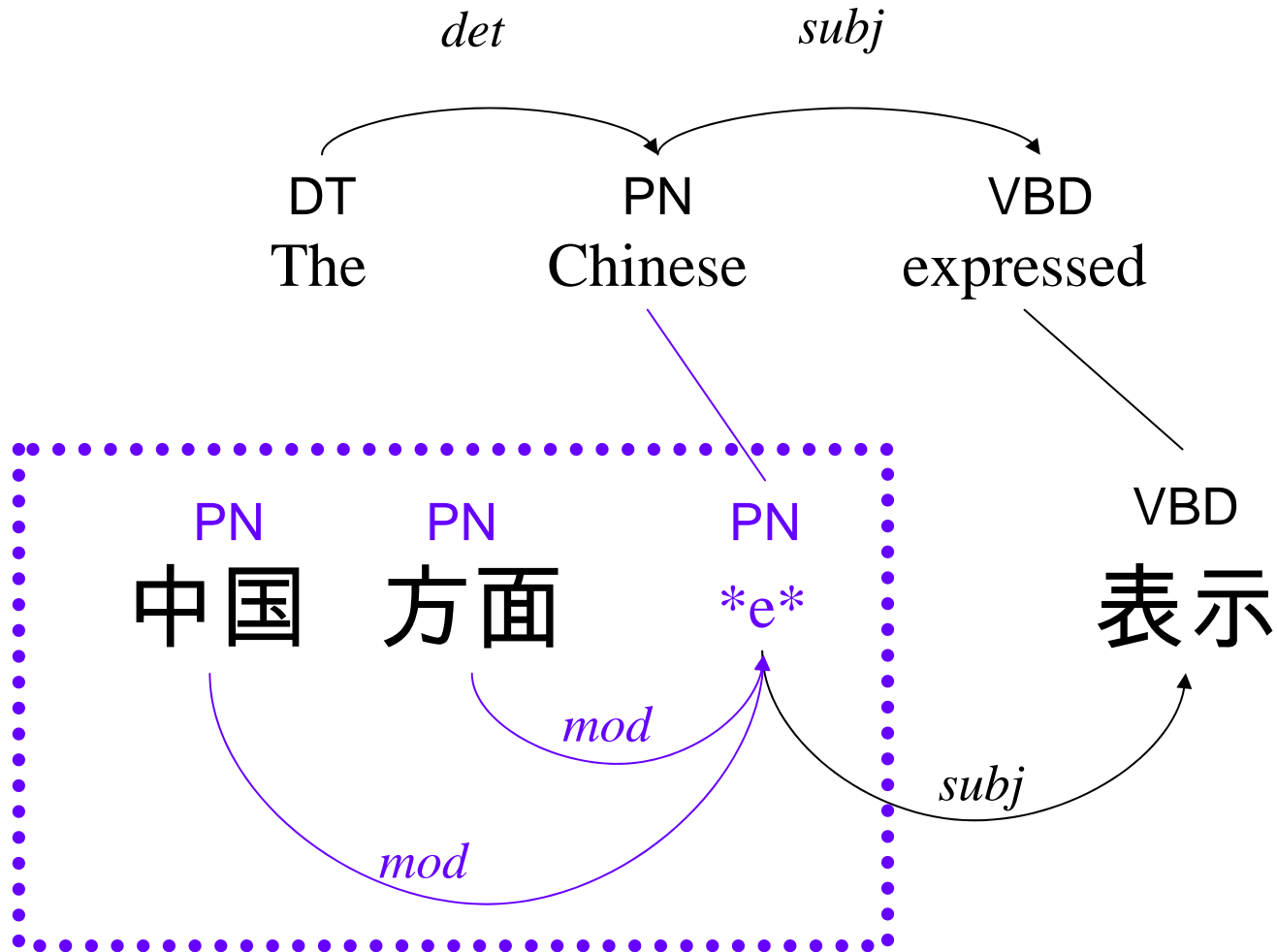
Many-to-1



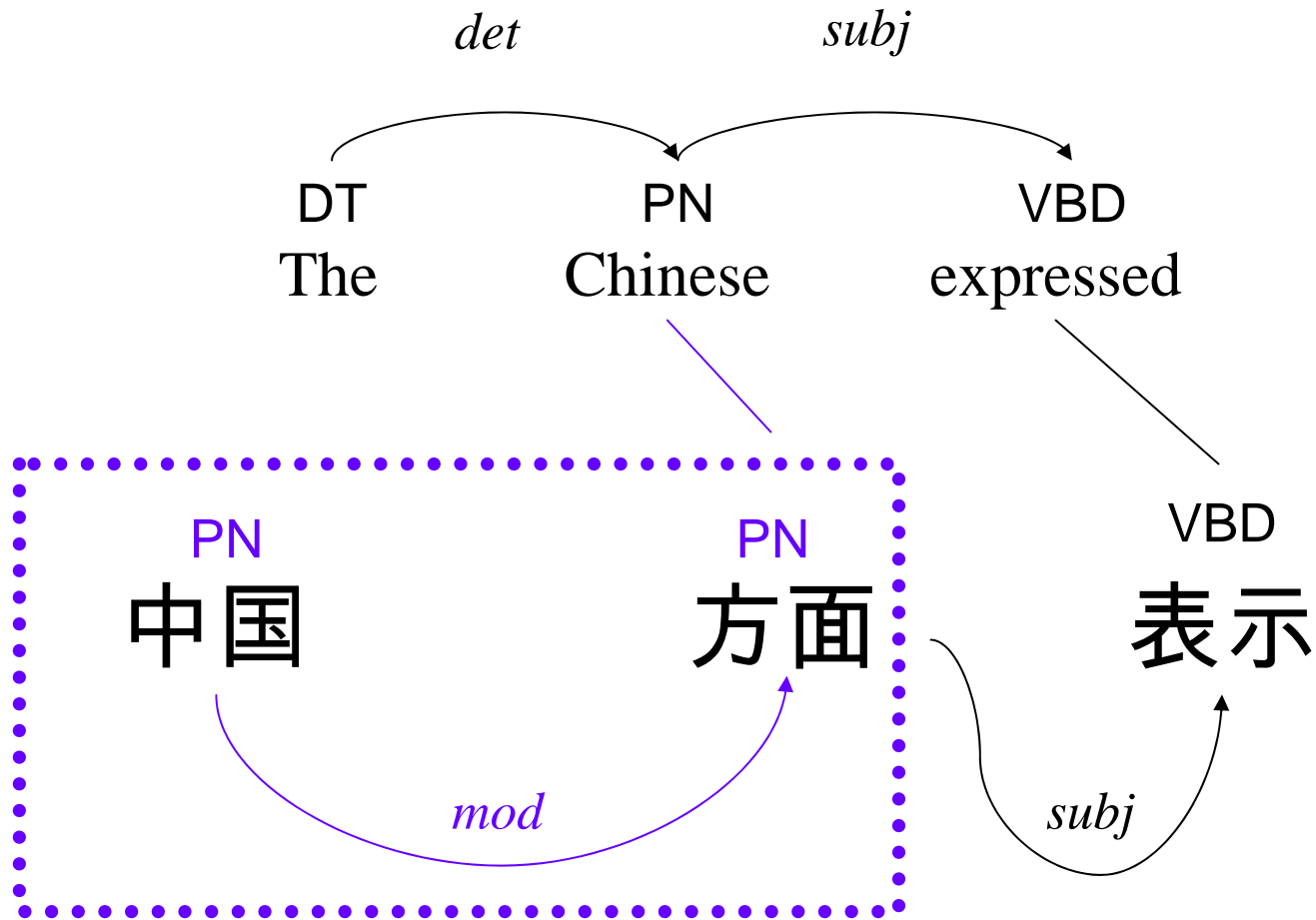
1-to-Many



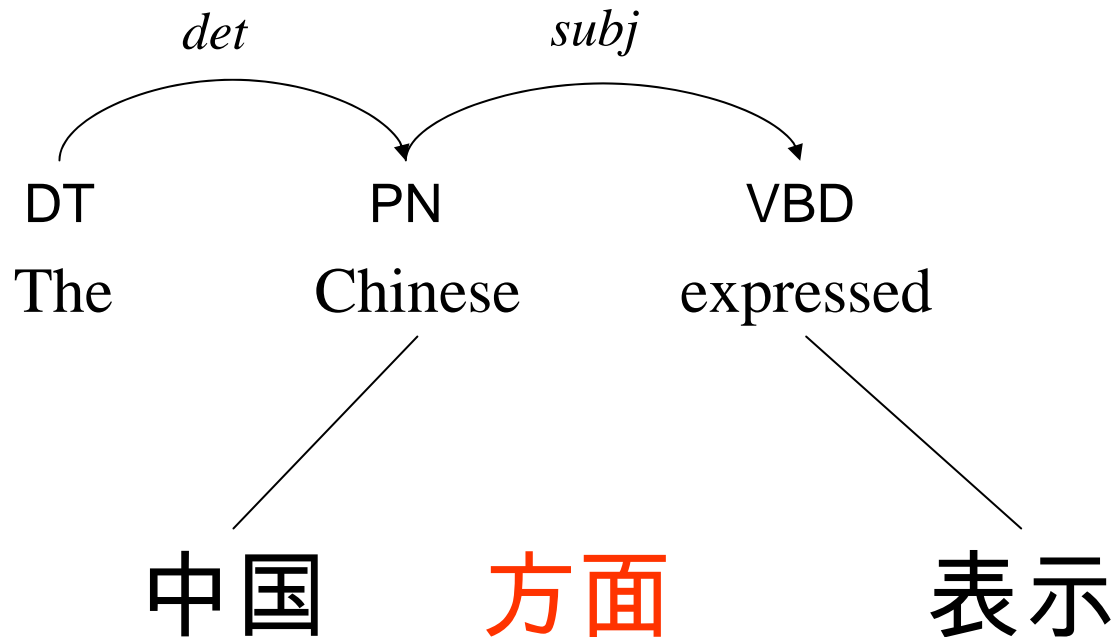
1-to-Many



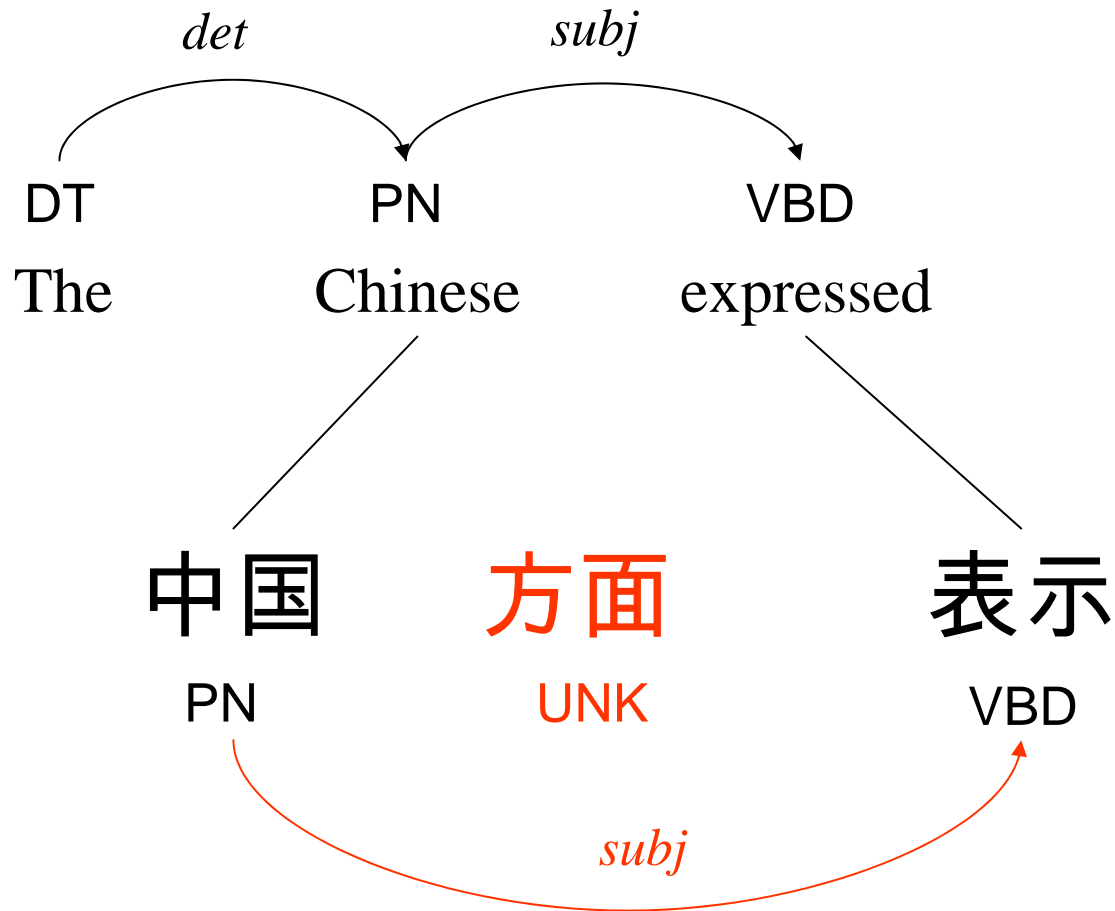
1-to-Many



Unaligned Chinese



Unaligned Chinese



Addressing Word Alignment Mismatches

- Use known linguistic facts about the target language to **transform the projected annotations**
- For **one-to-many** and **many-to-one** cases, select annotation based on grammatical categories
 - In Chinese, the head of a noun phrase is the last word
- Can incorporate some **unaligned words** back into the projected annotations
 - Functional words (e.g., aspectual, measure words)
 - Easily enumerable lexical categories (e.g., \$, RMB, yen)

Challenges: Component Errors

- English parser not perfect
 - Optimistically, 90% accurate (news text)
 - But only 86% on mixed genres like Brown
 - Still lower for imperfect English (e.g., blogs)
- Word alignments not perfect
 - For languages similar to English, low alignment error rate (10-15%)
 - For Chinese, alignment error rate is 40-50%

Improving Robustness against Component Errors

- **Filter** data that have been badly processed (e.g., suspected poor word alignments)
 - Too many words not aligned
 - Too many (non-consecutive) words of one lang. map to one word in the other lang.
- **Aggressive re-weighting** [Yarowsky & Ngai, 2001]
 - Re-normalize probability distributions to remove annotations with low likelihoods

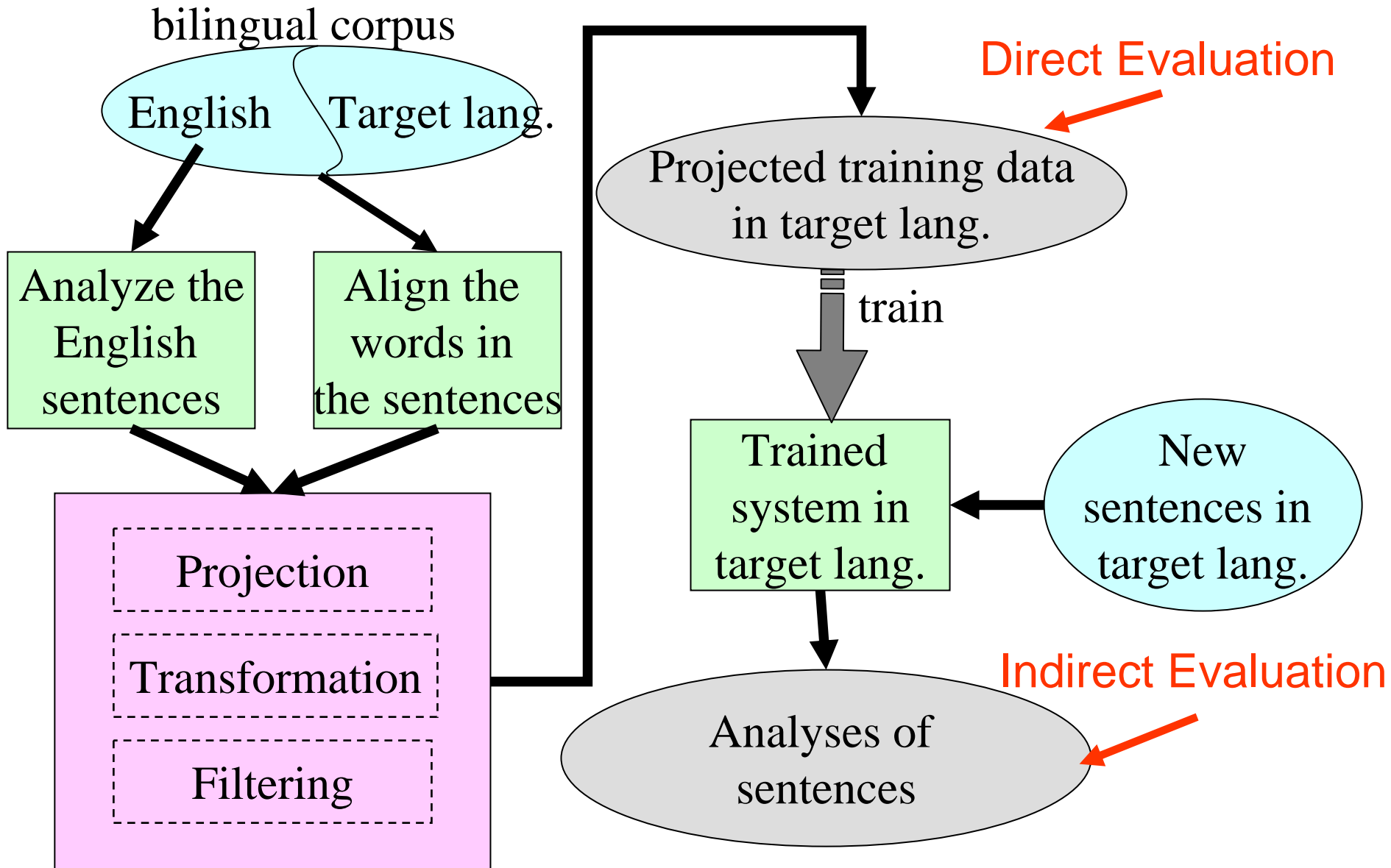
Roadmap

- Motivation
- Projected Annotation
- Empirical studies
 - Complexity of the learning tasks
 - Training a POS tagger and dependency parser
 - Similarities of the language pairs
 - English-French/Spanish
 - English-Chinese
 - Evaluation methods
- Future Work

Evaluation Methods

- **Direct evaluation:** how accurate are the projected resources?
 - Compare projected annotations against manually created gold standard directly
- **Indirect evaluation:** how accurate are the trained systems?
 - Use the projected resources to train a new system, then test it on new sentences and compare output against manually created gold standard
- Relationship between **performance degradation** and **component errors**
- Do **post projection transformation** and **filtering and re-weighting** help?

Projection Framework



Projected Tagging

- English-French
 - As reported by Yarowsky and Ngai (HLT-2001)
- English-Chinese
 - Data: 240K parallel sentences from FBIS
 - English Tagger: Ratnaparkhi's MaxEnt tagger
 - Alignment: IBM MT model 4 (GIZA++)
 - Test data: 1000 sentences from the ChTB (v4)
 - Chinese Tagger: Trigram Model

Direct Evaluation: Accuracy of Projected POS Tags

		English-French*	English-Chinese
Manually word aligned	Projection only	85%	63%
	+ Transformation		75%
Automatically word aligned	Projection only	76%	50%
	+ Transformation		65%

* English-French experiments are taken from Yarowsky and Ngai, *HLT 2001*

Indirect Evaluation: Accuracy of a Trained POS Tagger

	English-French*	English-Chinese
POS Tagger with supervised training	97%	93%
Baseline		53%
Projection only	86%	48%
+ Transformation		64%
+ Filtering and Re-weighting	96%	71%

* English-French experiments are taken from Yarowsky and Ngai, *HLT 2001*

Projected Tagging Findings

- More difficult to bootstrap a POS Tagger for Chinese than for French
 - English and French share similar tag sets
 - Fewer categorical divergences between English and French
 - Chinese has higher average number of tags per word than French
 - Re-weighting does not help as much for Chinese
- How can we close the gap?

Projected Parsing

- English Parser: Collins converted to dependency
- Alignment: IBM MT model 4 (GIZA++)
- English-Spanish
 - Parallel Data: 98K sentences from FBIS/Bible/UN
 - Test data: 200 unseen sentences from FBIS/Bible/UN
 - Upper bound parser: an off-the-shelf constraint grammar parser
- English-Chinese
 - Parallel Data: 240K sentences from FBIS
 - Test data: 2800 sentences from the ChTB (v4)

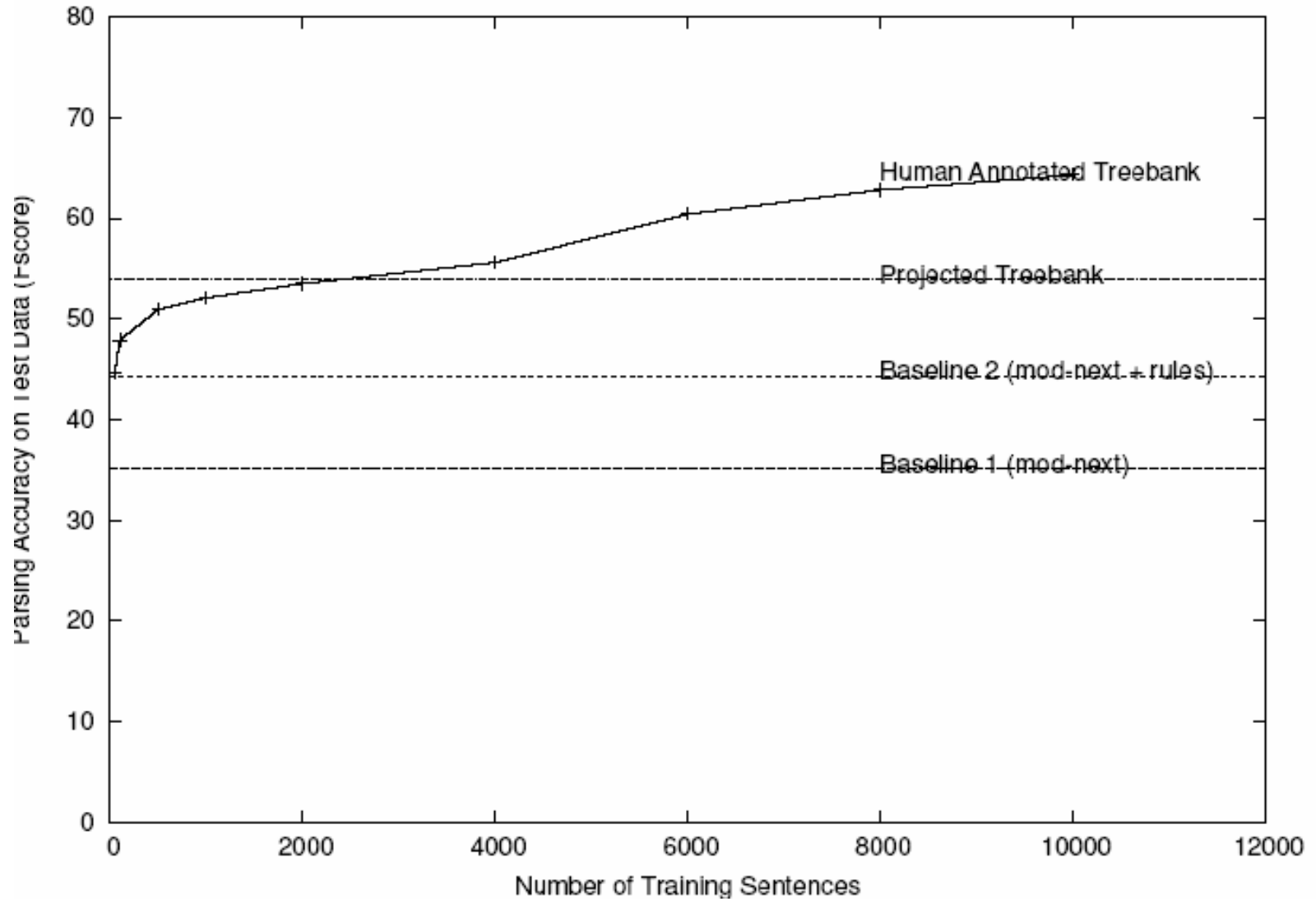
Direct Evaluation: Accuracy of Projected Parse Trees

		English-Spanish	English-Chinese
Manually word aligned	Projection only	37%	38%
	+ Transformation	70%	67%
Automatically word aligned	Projection only	34%	26%
	+ Transformation	66%	52%

Accuracy of a Trained Parser

	English-Spanish	English-Chinese
Resource-expensive parsers	69%	Chinese Treebank v4 64%**
Baseline (mod next)	34%	35%
+ Transformations	39%	44%
Projection		
+ Transformation	67%	
+ Filtering	72%	54%

Learning Curve Comparisons



Conclusion

- Use projection to acquire annotated resources for Chinese by bootstrapping from English resources
- The projected resources have an accuracy rate of nearly 70% in principle for both tagging and parsing.
- Reducing noise caused by word-alignment errors is still a major challenge.
- Systems trained on the induced resources outperform similarly inexpensive options.

Future Directions

- Reduce error rates of the word-alignment models
- Improve the projection algorithm to address more language divergences
- Develop more sophisticated techniques to filter out errors from the induced resources

Acknowledgements

Univ. of Maryland

Univ. of Pittsburgh

Philip Resnik

Chenhai Xi

Amy Weinberg

Karina Ivanetich

Clara Cabezas

Behrang Mohit

Okan Kolak

Carol Nichols

Adam Lopez

Reserve slides

English-French (Y&N table)

Model	Evaluate on E-F Aligned French		Evaluate on Unseen Monolingual French	
	Core Tagset	Eng Eqv Tagset	Core Tagset	Eng Eqv Tagset
(a) Direct transfer (auto-aligned, auto-project)	.76	.69	N/A	N/A
(b) Direct transfer (hand-aligned, auto-project)	.85	.78	N/A	N/A
(c) Standard bigram model (auto-aligned, auto-project)	.86	.82	.82	.68
(d) Noise-robust bigram induction (auto-aligned, auto-project)	.96	.93	.94	.91
(e) Standard bigram model (trained on heldout goldstandard)	.97	.96	.98	.97

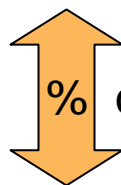
Table 4: Evaluation of 5 POS tagger induction models on 2 French datasets and 2 tagsets

Next Step: Filtering:

Explore what training data leads to better accuracy



% of English, no Chinese
in each sentence



% of Chinese, no English
in each sentence

1 To None: Chinese, no English	20.1% (97% of OTHER)
1 To None: English, no Chinese	30.9%
1 Chinese to Many English	14.4%
1 Chinese to 1 English	34.6%

Why Filtering May Improve
Accuracy:

Percentages of
Correspondences over All
Words

Initial Accuracy Results

	Test on ChTB (1165)	Test on ChTB Core (1165)
Train on ChTB (14,000)	92.7%	
Train on ChTB Core (14,000)		92.9%
Train on FBIS Core (240,000)		48.2%

Accuracy after Filtering

C no E / E no C	.4	.3	.2	.1
.4	48.0 183,714	49.6 139,722	51.5 76,605	53.9 17768
.3	48.0 162,831	49.4 122,302	51.06 64,310	53.6 13265
.2	47.1 81,981	48.7 55,672	51.1 24,952	52.8 4,312
.1	45.2 7417	46.4 4683	48.7 2,180	46.1 729

blue = % accuracy

green = number of sentences

Accuracy after Filtering

	Chinese-no-English			
English -no-Chinese	.4	.3	.2	.1
.4	64.0 183,714	66.1 139,722	67.7 76,605	69.9 17,768
.3	64.0 162,831	66.2 122,302	67.7 64,310	70.0 13,265
.2	63.5 81,981	65.6 55,672	67.7 24,952	70.6 4,312
.1	63.1 7417	64.4 4683	66.7 2,180	65.7 729

Next Step: Will Re-Weighting POS tags Improve Accuracy?

Average Number of Tags n by Word Frequency

Overall	$n \geq 100$	$5 < n < 100$	$n \leq 5$
3.8	6.7	4.3	2.0

% of Occurrence of Most Frequent Tags by Word Frequency

Overall	$n \geq 100$	$5 < n < 100$	$n \leq 5$
59%	56%	55%	66%