Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents

Dun Deng, **Nianwen Xue**, Shiman Guo

Brandeis University

2015-06-04

Denver, Colorado

What’s an ideal language resource for statistical machine translation?

- Lots and lots of parallel texts?
- A large word and phrase aligned parallel treebank?
- A large semantically annotated parallel corpus?
  - Parallel Abstract Meaning Representation Corpus
- Other

What’s an ideal language resource for statistical machine translation?

- Lots and lots of parallel texts?

✓ **A large word and phrase aligned parallel treebank?**

- A large semantically annotated parallel corpus?
  - Parallel Abstract Meaning Representation Corpus
- Other

Issues with treebanks and word alignments

- Status quo:
  - Manually annotated treebanks and word alignments independently conceived
    - Inevitable incompatibilities between word alignment and the syntactic structure
  - Treebanks on both sides are independently conceived
    - Incompatibilities between the parse trees of a sentence pair
  - Treebanks not optimized for MT
    - Trees too shallow, too deep
- Frustrated MT users
  - Forget about parse trees!
  - Or use automatically induced parse trees!

Addressing this problem

- Create a hierarchically aligned Chinese English parallel treebank that
  - harmonizes word alignment and the syntactic structure of a sentence,
  - synchronizes the parse trees of the sentence pair
  - Empirically determine the amount of structure that is needed during alignment

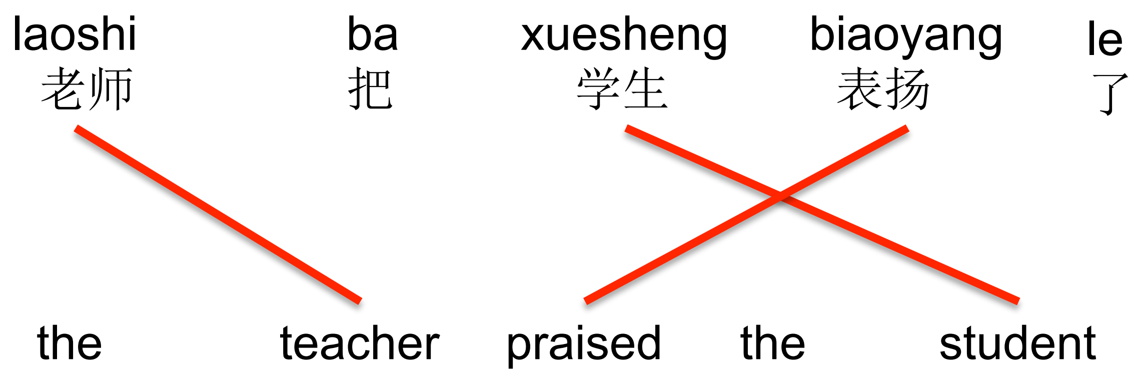
Word alignment is deceptively complicated

✓ An equivalent exists in the target-language sentence, which matches the word in both lexical meaning and grammatical function

? There is a candidate in the target-language sentence, which does not have the same lexical meaning and/or grammatical function as the word but could be aligned in the given context

? The word has no translation counterpart in the target-language sentence at all.

Word alignment

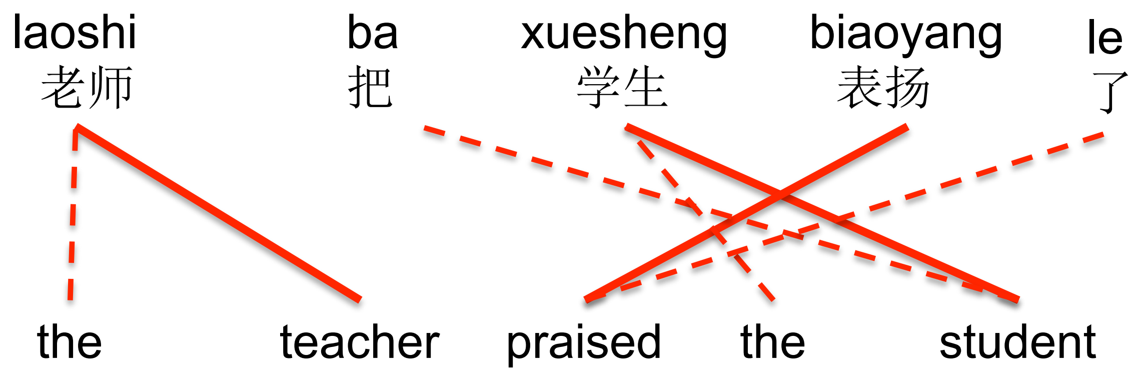


laoshi ba xuesheng biaoyang le

老师 把 学生 表扬 了

the teacher praised the student

Word alignment



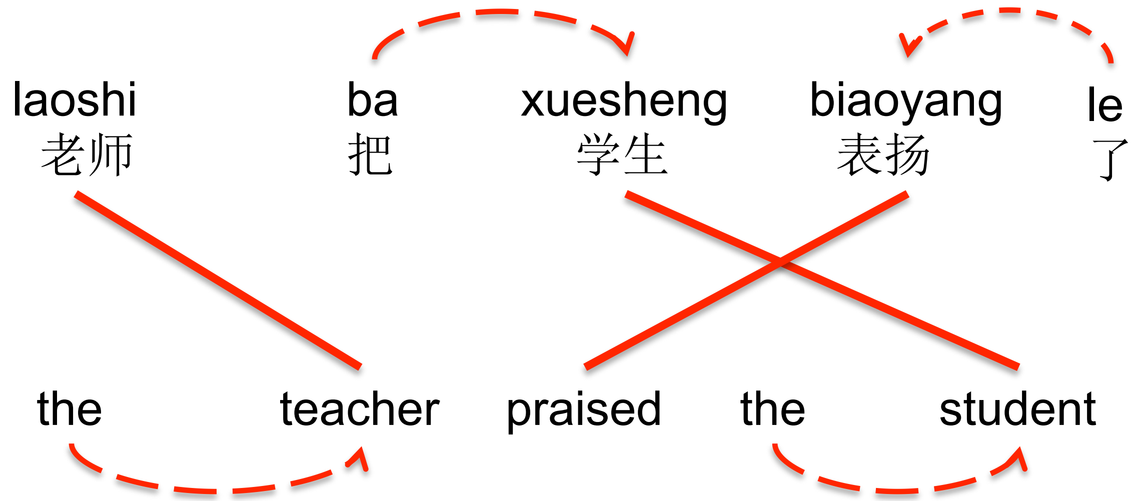
laoshi 老师 ba 把 xuesheng 学生 biaoyang 表扬 le 了

the teacher praised the student

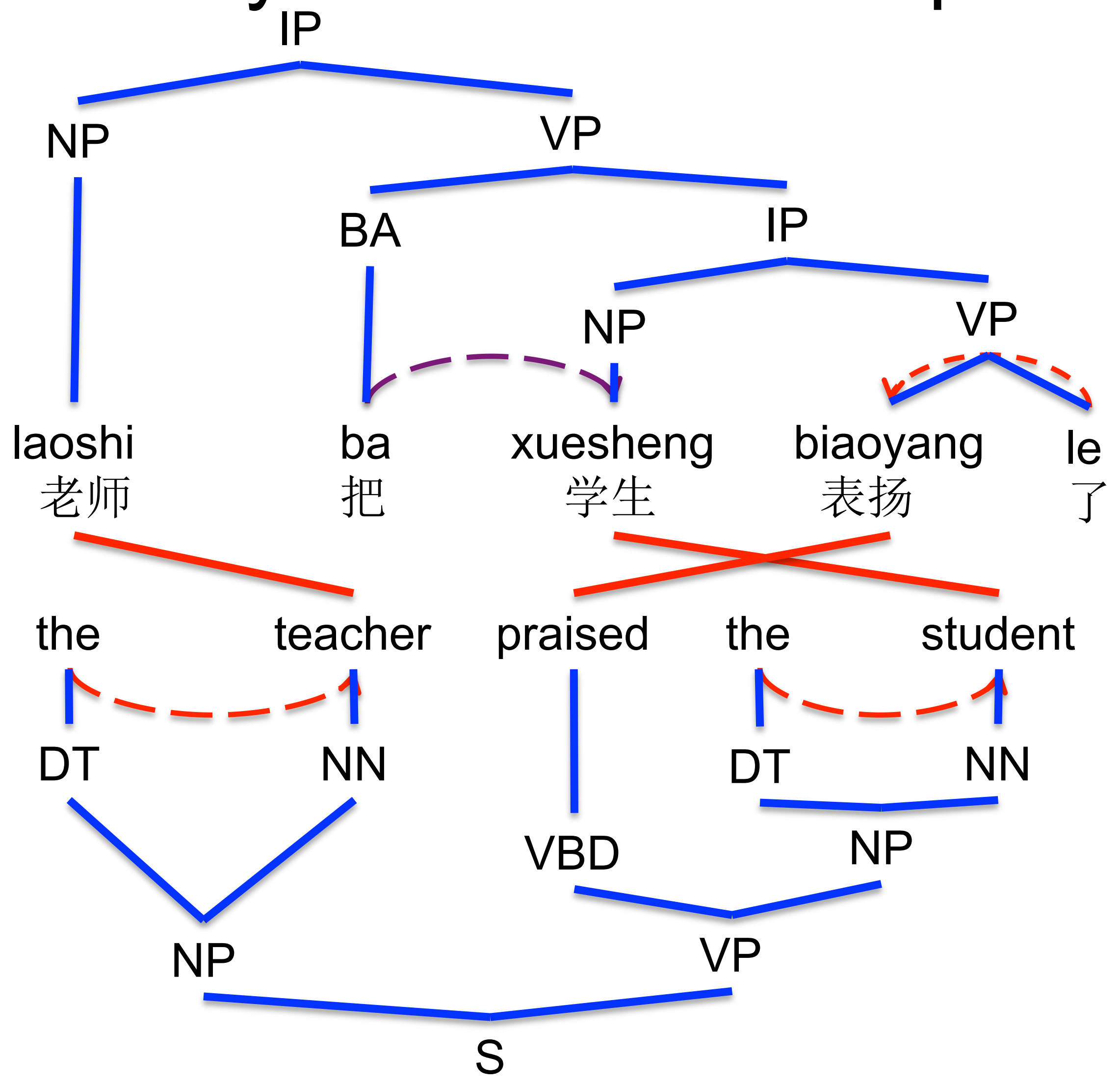
老师⇔teacher
老师⇔the teacher

学生⇔student
学生⇔the student
把学生⇔student
把学生⇔the student

“Syntactic annotation” in word alignment



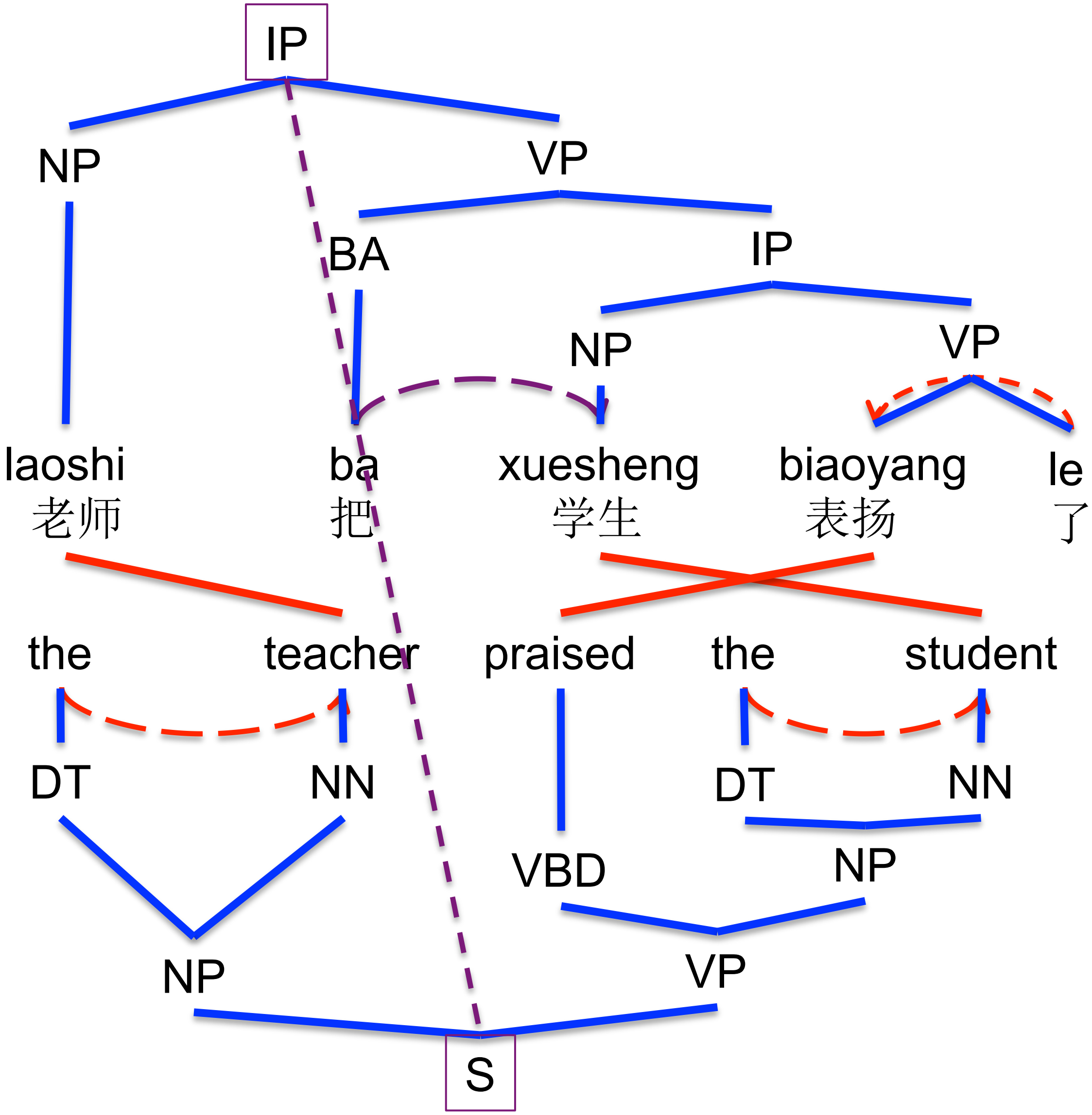
Redundancy and conflict with parse trees



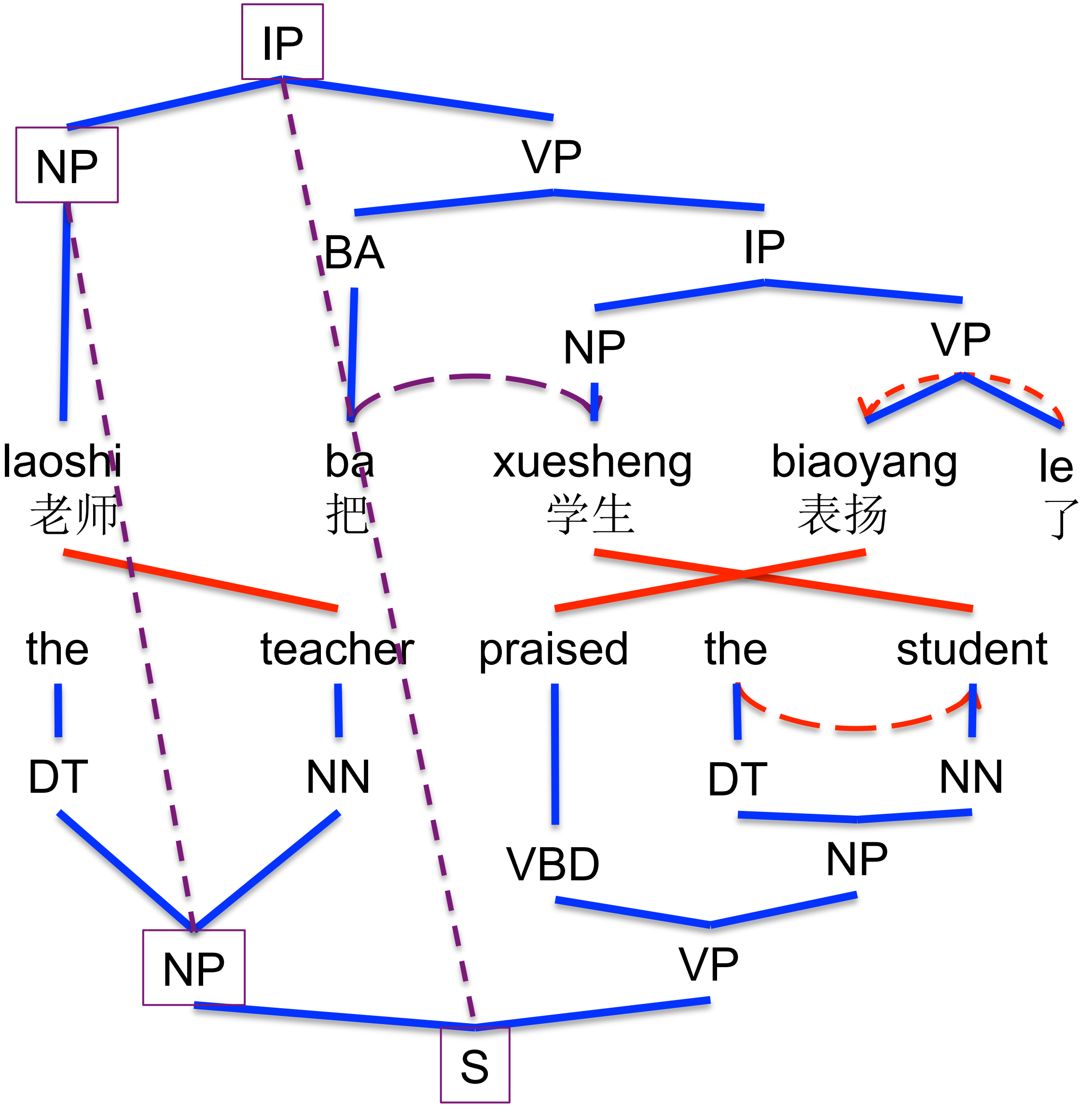
Causes of conflict

- Word alignment standards and treebanking standards are independently established and are meant to be used separately
- Need an approach that systematically considers the interaction between word alignment and syntactic structure to maximize the utility of aligned parse trees
- Solution: hierarchical alignment
  - “sure” alignments only at word level, others aligned indirectly via node alignment in their context

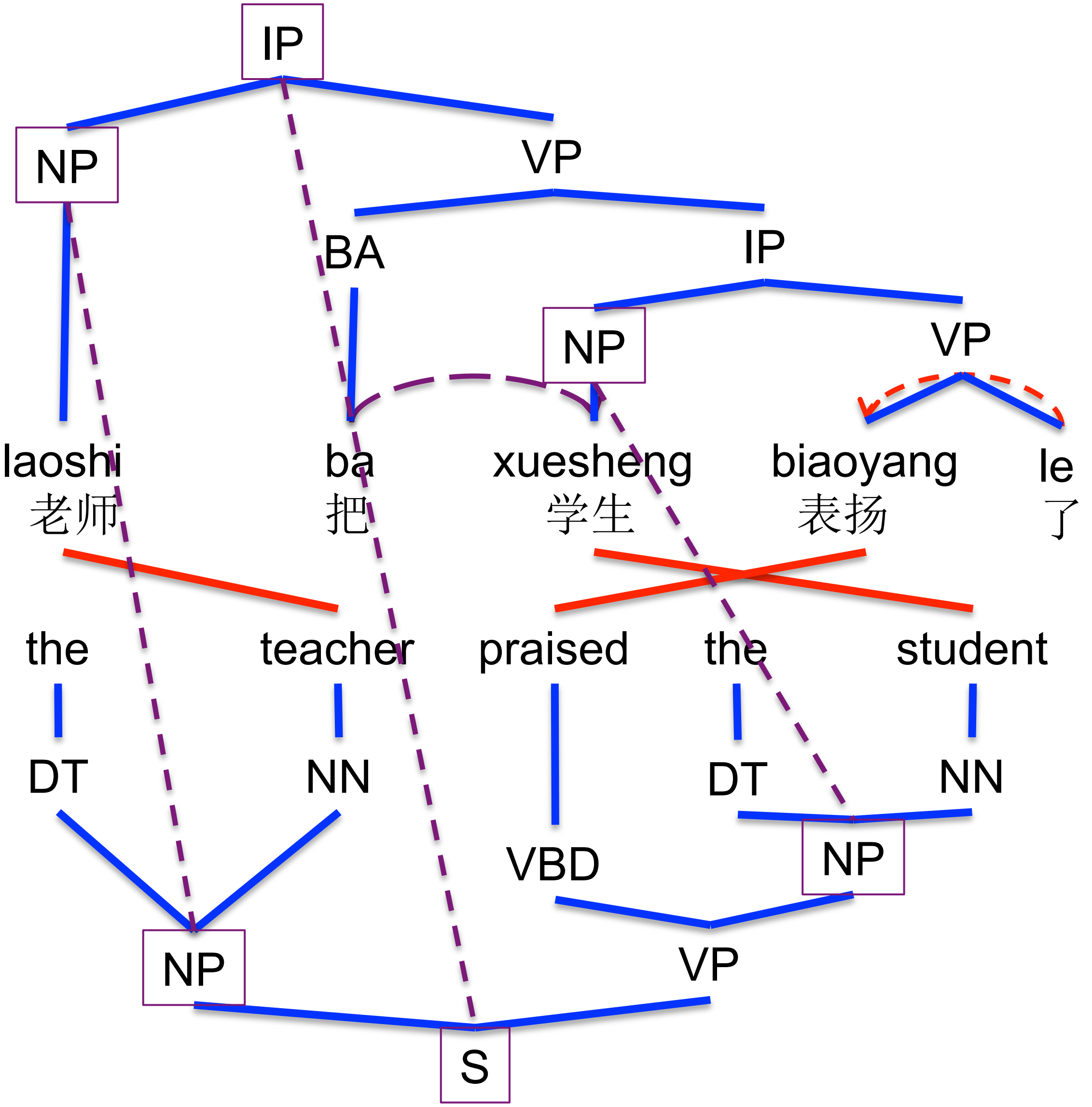
Hierarchical alignment



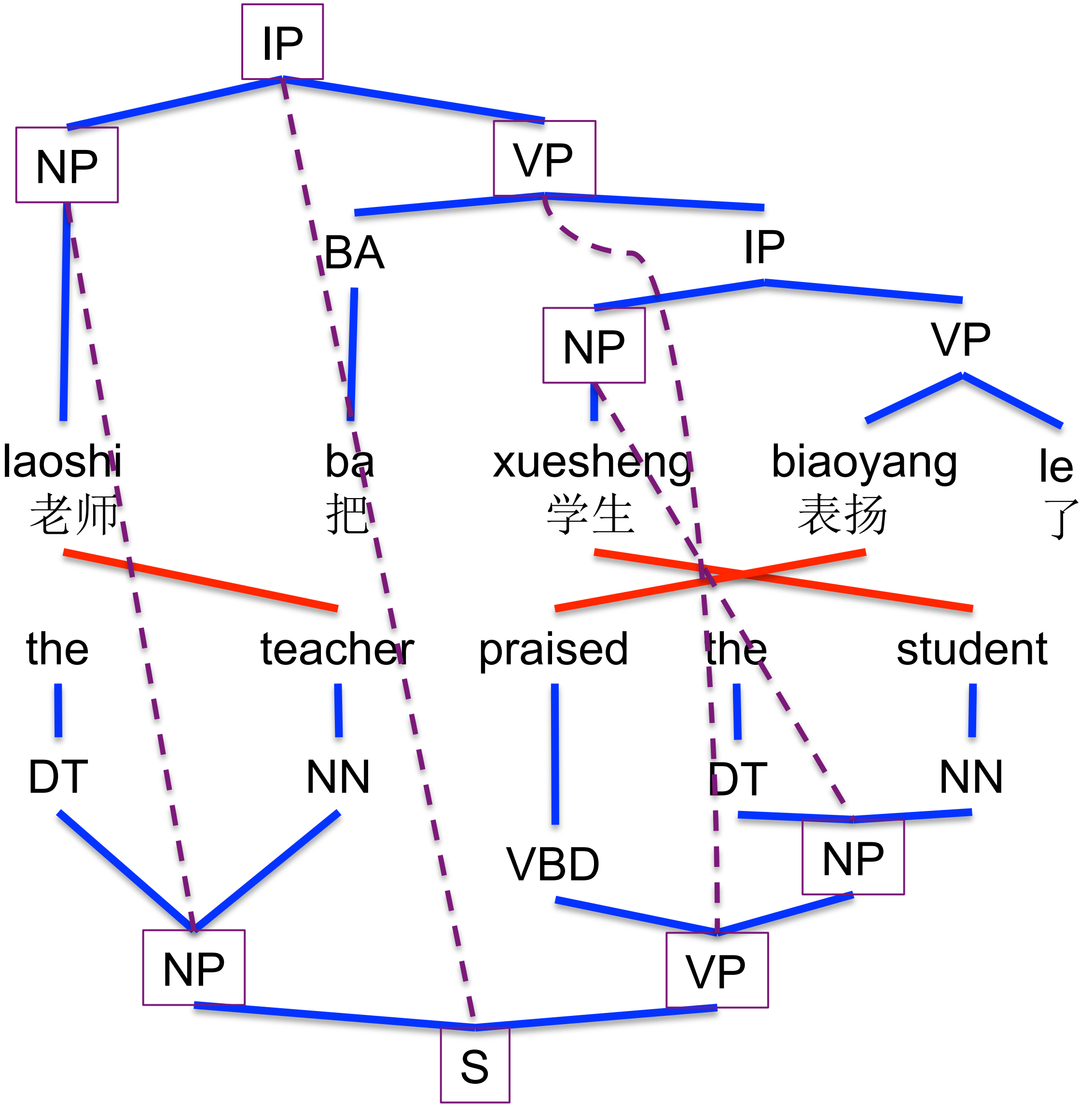
Hierarchical alignment



Hierarchical alignment



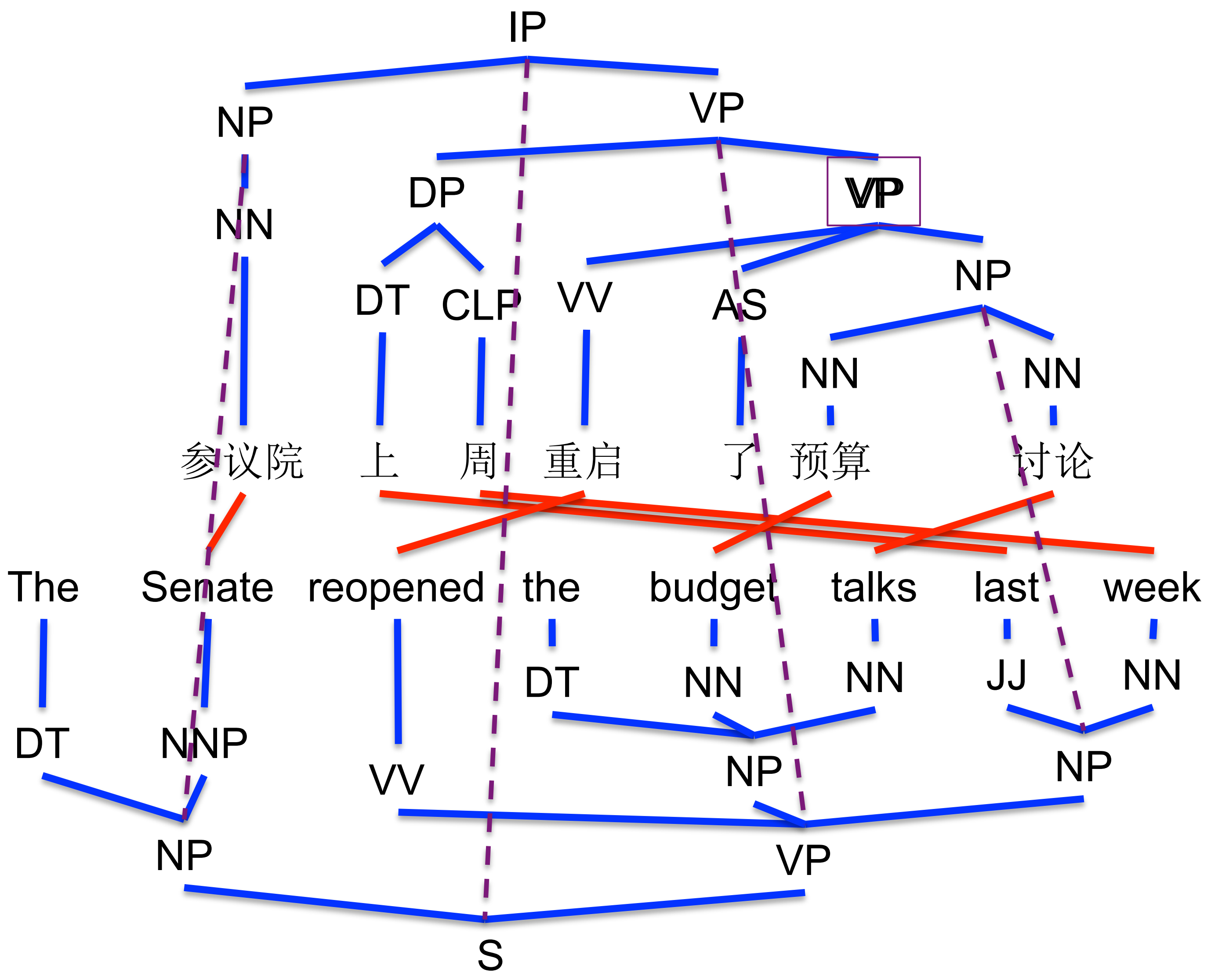
Hierarchical alignment



Issues in aligning parse trees

- Too much structure
  - Not all nodes in a parse tree are aligned or alignable.
  - This might suggest that not all the syntactic structures annotated in existing treebanks are necessary for MT purposes
- Too little structure
  - Flat structures in the parse trees can prevent legitimate alignments
  - Needs to revise trees in order to get proper alignments

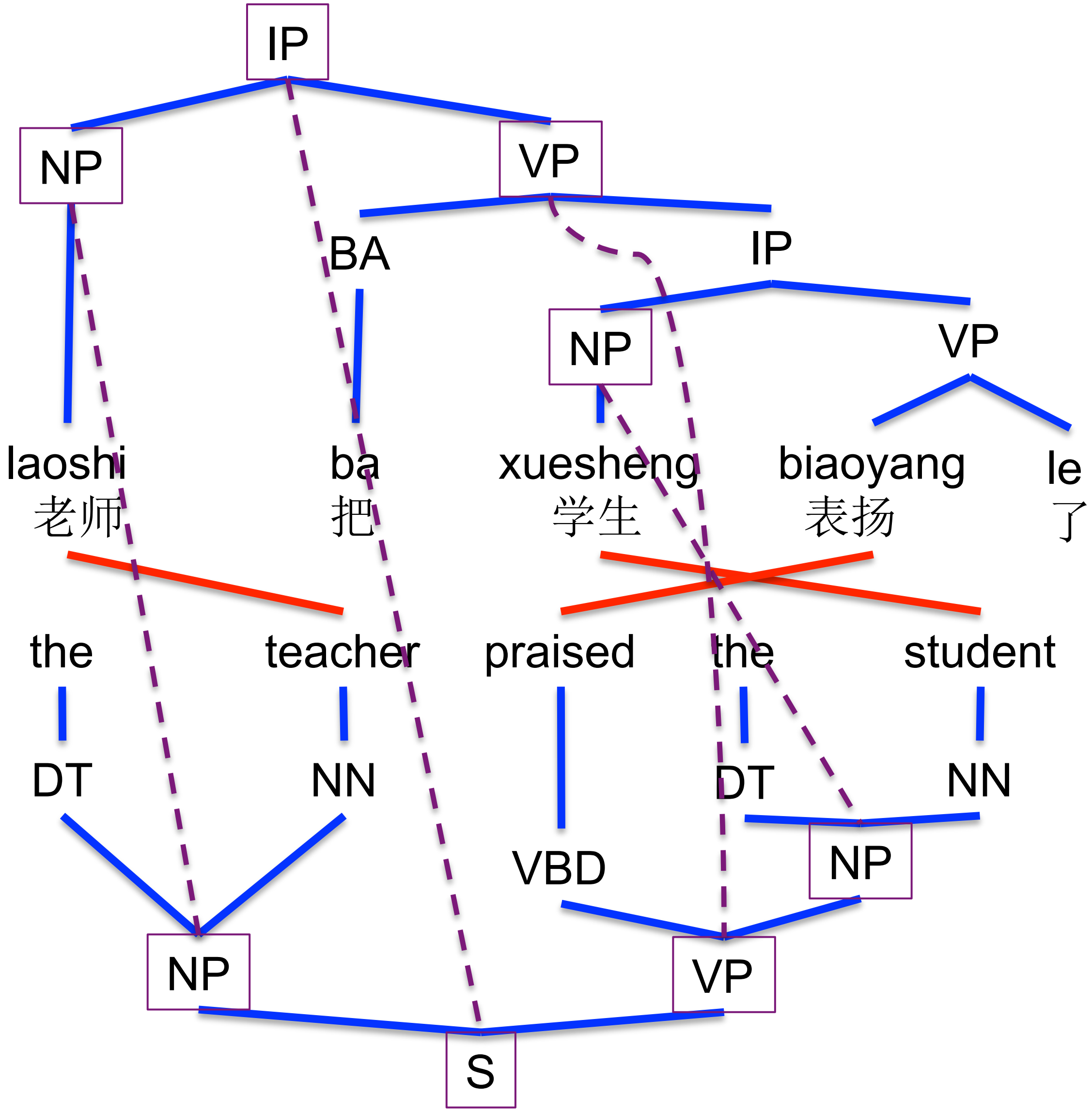
Too little structure



Potential benefits of hierarchical alignment

- Captures unconstrained long-distance lexical dependencies
- Allows the extraction of linguistically interpretable rules
- Smaller but equally potent translation table
- Better MT accuracy?

Fewer rules



学生⇔the student

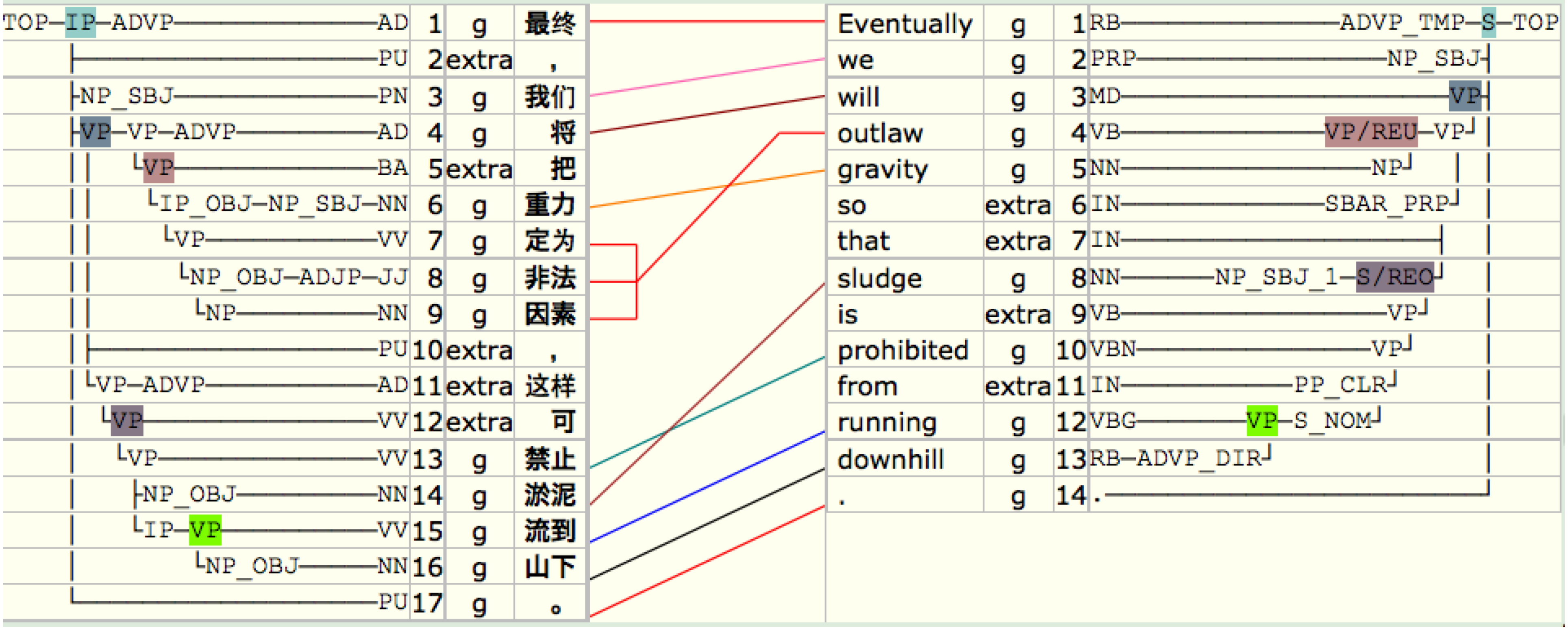
~~把学生⇔the stuent~~

~~把学生⇔student~~

把X表扬了⇔praised X

Long distance dependency: Passives

S66



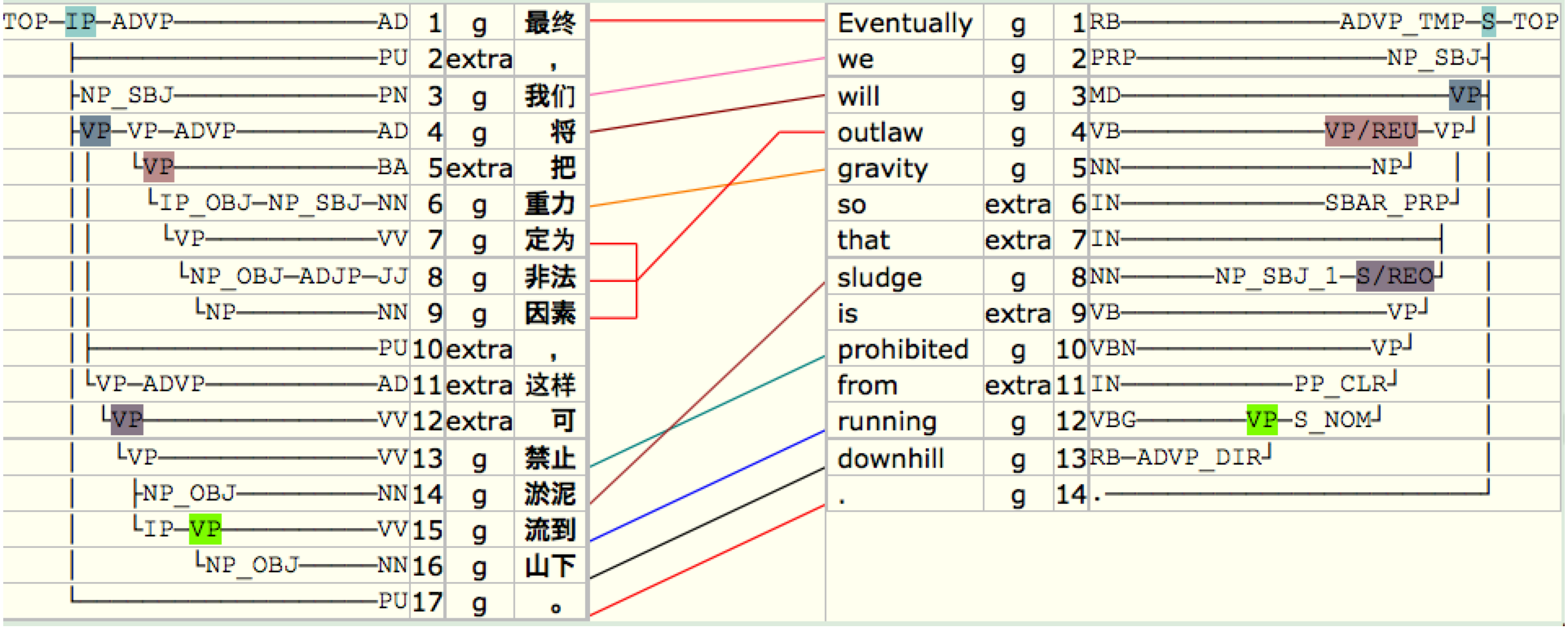
TOP—IP—ADVP——AD	1	g	最终
├——PU	2	extra	，
├NP_SBJ——PN	3	g	我们
├VP—VP—ADVP——AD	4	g	将
└VP——BA	5	extra	把
└IP_OBJ—NP_SBJ—NN	6	g	重力
└VP——VV	7	g	定为
└NP_OBJ—ADJP—JJ	8	g	非法
└NP——NN	9	g	因素
├——PU	10	extra	，
└VP—ADVP——AD	11	extra	这样
└VP——VV	12	extra	可
└VP——VV	13	g	禁止
├NP_OBJ——NN	14	g	淤泥
└IP—VP——VV	15	g	流到
└NP_OBJ——NN	16	g	山下
└——PU	17	g	。

Eventually	g	1	RB——ADVP_TMP—S—TOP
we	g	2	PRP——NP_SBJ┤
will	g	3	MD——VP┤
outlaw	g	4	VB——VP/REU—VP┘
gravity	g	5	NN——NP┘
so	extra	6	IN——SBAR_PRP┘
that	extra	7	IN——┤
sludge	g	8	NN——NP_SBJ_1—S/REO┘
is	extra	9	VB——VP┘
prohibited	g	10	VBN——VP┘
from	extra	11	IN——PP_CLR┘
running	g	12	VBG——VP—S_NOM┘
downhill	g	13	RB—ADVP_DIR┘
.	g	14	.——┘

[可 禁止 X_1 X_2 <--> X_1 is prohibited from X_2]

Long distance dependency: 把 construction

S66



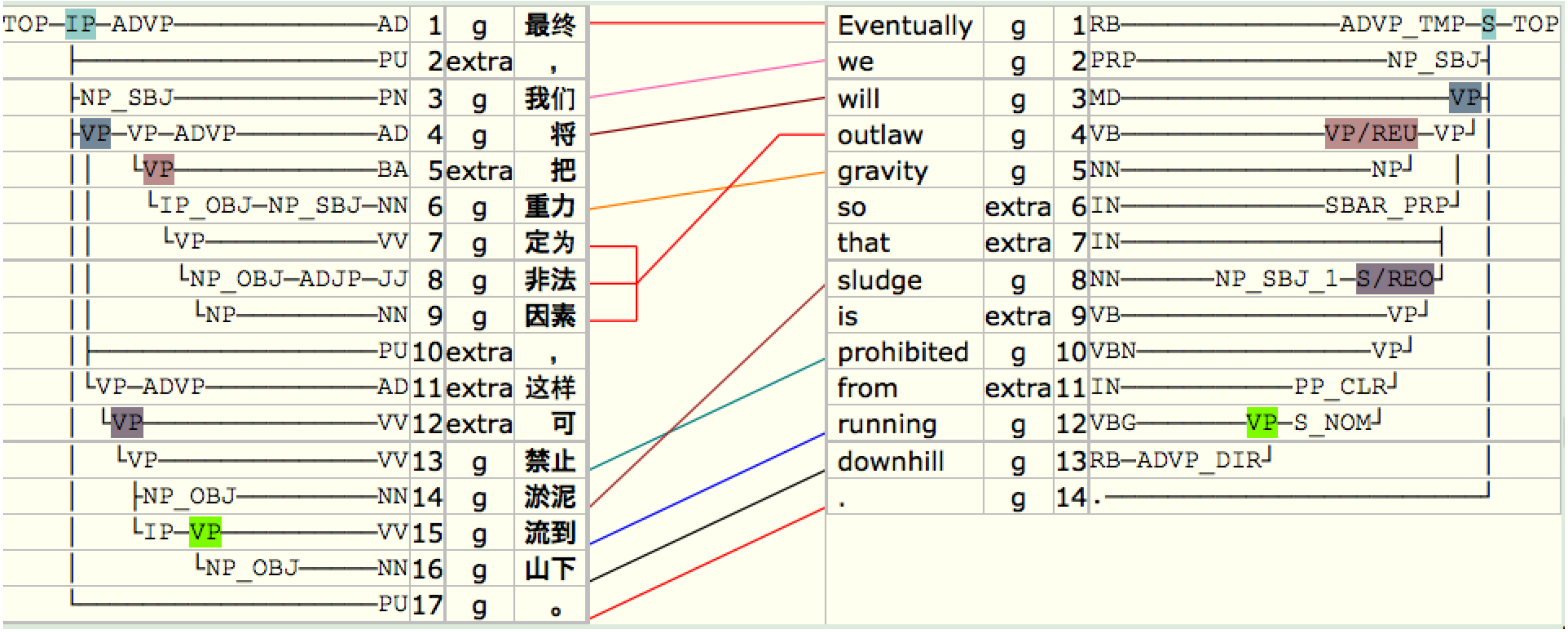
TOP─IP─ADVP─AD	1	g	最终
├─PU	2	extra	，
├NP_SBJ─PN	3	g	我们
├VP─VP─ADVP─AD	4	g	将
└VP─BA	5	extra	把
└IP_OBJ─NP_SBJ─NN	6	g	重力
└VP─VV	7	g	定为
└NP_OBJ─ADJP─JJ	8	g	非法
└NP─NN	9	g	因素
├─PU	10	extra	，
└VP─ADVP─AD	11	extra	这样
└VP─VV	12	extra	可
└VP─VV	13	g	禁止
├NP_OBJ─NN	14	g	淤泥
└IP─VP─VV	15	g	流到
└NP_OBJ─NN	16	g	山下
└─PU	17	g	。

Eventually	g	1	RB─ADVP_TMP─S─TOP
we	g	2	PRP─NP_SBJ┤
will	g	3	MD─VP┤
outlaw	g	4	VB─VP/REU─VP┘
gravity	g	5	NN─NP┘
so	extra	6	IN─SBAR_PRP┘
that	extra	7	IN─┤
sludge	g	8	NN─NP_SBJ_1─S/REO┘
is	extra	9	VB─VP┘
prohibited	g	10	VBN─VP┘
from	extra	11	IN─PP_CLR┘
running	g	12	VBG─VP─S_NOM┘
downhill	g	13	RB─ADVP_DIR┘
.	g	14	.─┘

[把 X 定为非法因素 <--> outlaw X]

Long distance dependency: discourse Connectives

S66



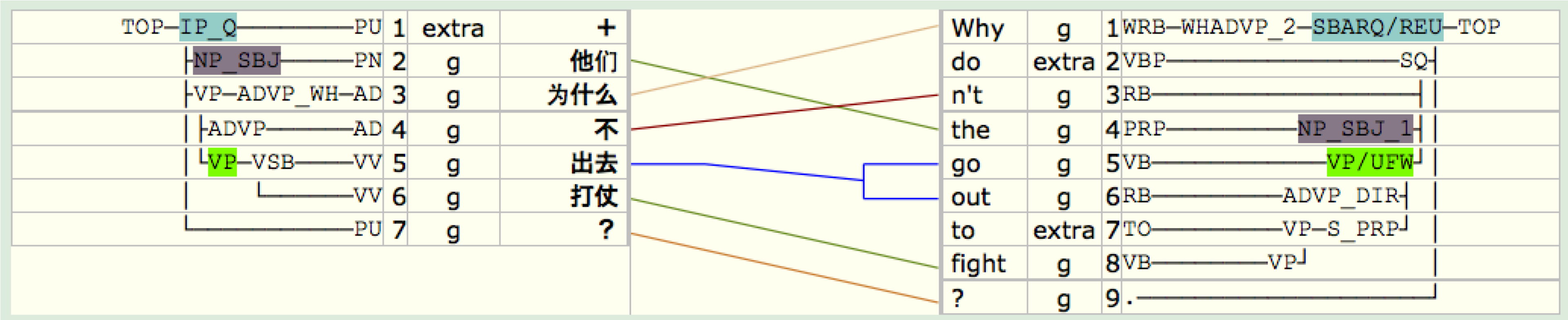
TOP—IP—ADVP——AD	1	g	最终
├——PU	2	extra	，
├NP_SBJ——PN	3	g	我们
├VP—VP—ADVP——AD	4	g	将
└VP——BA	5	extra	把
└IP_OBJ—NP_SBJ—NN	6	g	重力
└VP——VV	7	g	定为
└NP_OBJ—ADJP—JJ	8	g	非法
└NP——NN	9	g	因素
├——PU	10	extra	，
└VP—ADVP——AD	11	extra	这样
└VP——VV	12	extra	可
└VP——VV	13	g	禁止
├NP_OBJ——NN	14	g	淤泥
└IP—VP——VV	15	g	流到
└NP_OBJ——NN	16	g	山下
└——PU	17	g	。

Eventually	g	1	RB——ADVP_TMP—S—TOP
we	g	2	PRP——NP_SBJ┤
will	g	3	MD——VP┤
outlaw	g	4	VB——VP/REU—VP┘
gravity	g	5	NN——NP┘
so	extra	6	IN——SBAR_PRP┘
that	extra	7	IN——
sludge	g	8	NN——NP_SBJ_1—S/REO┘
is	extra	9	VB——VP┘
prohibited	g	10	VBN——VP┘
from	extra	11	IN——PP_CLR┘
running	g	12	VBG——VP—S_NOM┘
downhill	g	13	RB—ADVP_DIR┘
.	g	14	.

[将 X_1 ， 这样 X_2 <--> will X_1 so that X_2]

Long distance dependency: Questions

S1550



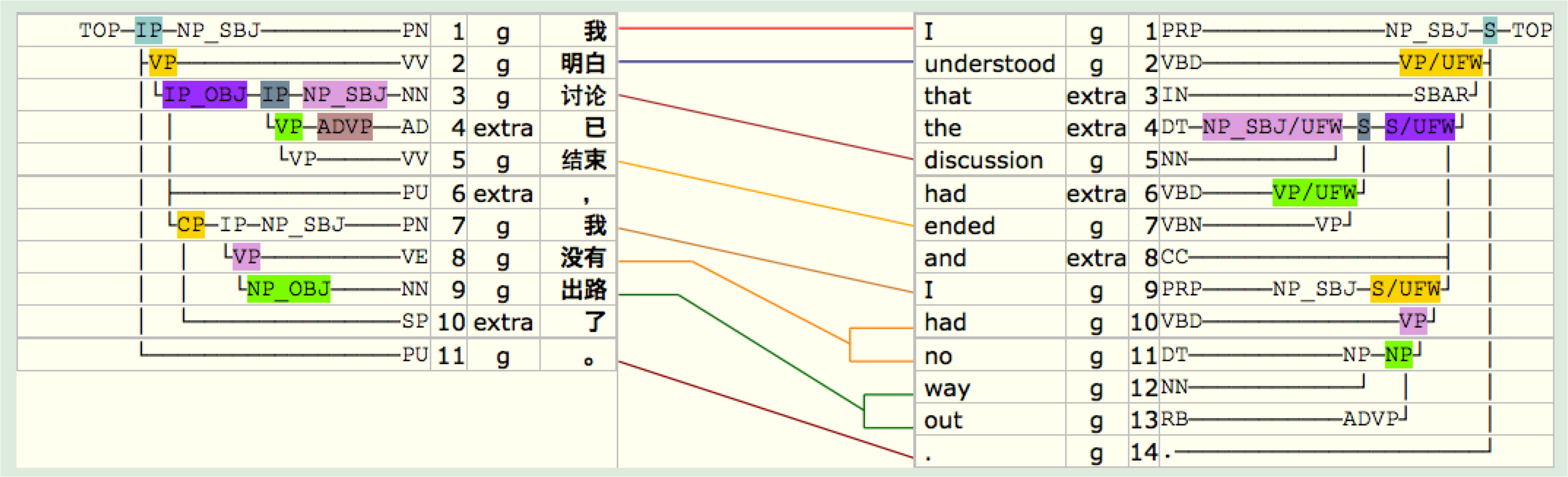
TOP─IP_Q─────PU	1	extra	+
├NP_SBJ────PN	2	g	他们
├VP─ADVP_WH─AD	3	g	为什么
│ ├ADVP────AD	4	g	不
│ └VP─VSB───VV	5	g	出去
│ └────VV	6	g	打仗
└────────PU	7	g	？

Why	g	1	WRB─WHADVP_2─SBARQ/REU─TOP
do	extra	2	VBP──────SQ┤
n't	g	3	RB──────┤ │
the	g	4	PRP────NP_SBJ_1┤ │
go	g	5	VB─────VP/UFW┘ │
out	g	6	RB────ADVP_DIR┤ │
to	extra	7	TO────VP─S_PRP┘ │
fight	g	8	VB────VP┘ │
?	g	9	.────────┘

[X_1 为什么 不 X_2 <--> Why do n’t X_1 X_2]

Alignment in syntactic context: Conjunction

S3850



PN	1	g	我
VV	2	g	明白
NN	3	g	讨论
AD	4	extra	已
VV	5	g	结束
PU	6	extra	，
PN	7	g	我
VE	8	g	没有
NN	9	g	出路
SP	10	extra	了
PU	11	g	。

I	g	1	PRP
understood	g	2	VBD
that	extra	3	IN
the	extra	4	DT
discussion	g	5	NN
had	extra	6	VBD
ended	g	7	VBN
and	extra	8	CC
I	g	9	PRP
had	g	10	VBD
no	g	11	DT
way	g	12	NN
out	g	13	RB
.	g	14	.

[X , Y <> X and Y]

Conclusions and future work

- Described a hierarchical alignment approach that systematically considers the interaction between word alignment and the syntactic structure of a sentence.
- Showed that such alignments can be used to extract translation rules that cover long-distance dependencies.
- Aligned 10K sentence pairs annotated with PTB and CTB trees
- To do: revising treebanking guidelines and synchronizing the parse trees
- More info can be found at:
  - http://www.cs.brandeis.edu/~clp/ctb/hacept/

Acknowledgements

- Libin Shen for inspirations
- Tool support by the IBM Multilingual NLP Technologies Group: Niyu Ge, Abe Ittycheriah, Salim Roukos
- Annotators: Gao Hui, Guo Shiman, Wang Tse-Ming, Zhou Linya
- Funded by DARPA BOLT Program via IBM