Supervised Treebank Conversion: Data and Approaches (ACL-2018)



Xinzhou Jiang[†], Bo Zhang[†], Zhenghua Li[†], Min Zhang[†], Sheng Li[‡], Luo Si[‡],

[†] Soochow University, Suzhou, China; [‡] Alibaba Inc., Hangzhou, China

zhli13@suda.edu.cn; http://hlt.suda.edu.cn/~zhli



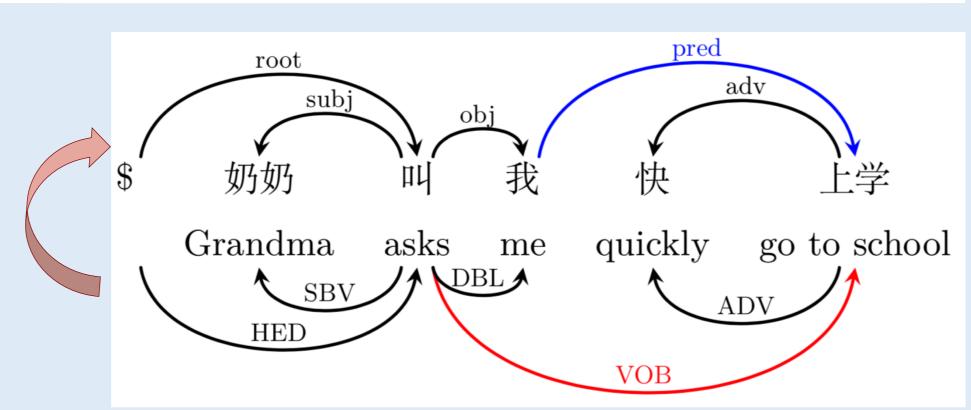
Motivation

- Many languages have multi heterogeneous treebanks, with new ones being annotated all the time.
- To meet the needs of certain applications, or
- Motivated by different linguistic philosophies

Questions

- Can we accurately convert a guideline-A tree to a guideline-B tree? (the conversion task)
- Can we improve parsing accuracy by enlarging training data with converted data? (the parsing task)

	// 70 - 1	
Treebanks	#Tok	Grammar
Sinica (Chen et al., 2003)	0.36M	Case grammar
CTB (Xue et al., 2005)	1.62M	Phrase structure
TCT (Zhou, 2004)	1.00M	Phrase structure
PCT (Zhan, 2012)	0.90M	Phrase structure
HIT-CDT (Che et al., 2012)	0.90M	Dependency structure
PKU-CDT (Qiu et al., 2014)	1.40M	Dependency structure



Previous works on multi-treebank exploitation

- The indirect guide-feature method (stacked learning) (Li+, ACL-2012)
- The indirect multi-task learning (MTL) method (Guo+, COLING-2016)
- The direct treebank conversion method
 - Unsupervised (Niu+, ACL-2009)
- Using pseudo bi-tree aligned data (Zhu+ ACL-2011; Li+ ACL-2013)

Contributions of this work

- First propose the task of supervised treebank conversion
- Manually annotate a bi-tree aligned dataset for training and evaluation
- Propose two simple yet effective approaches

Annotation of bi-tree aligned data

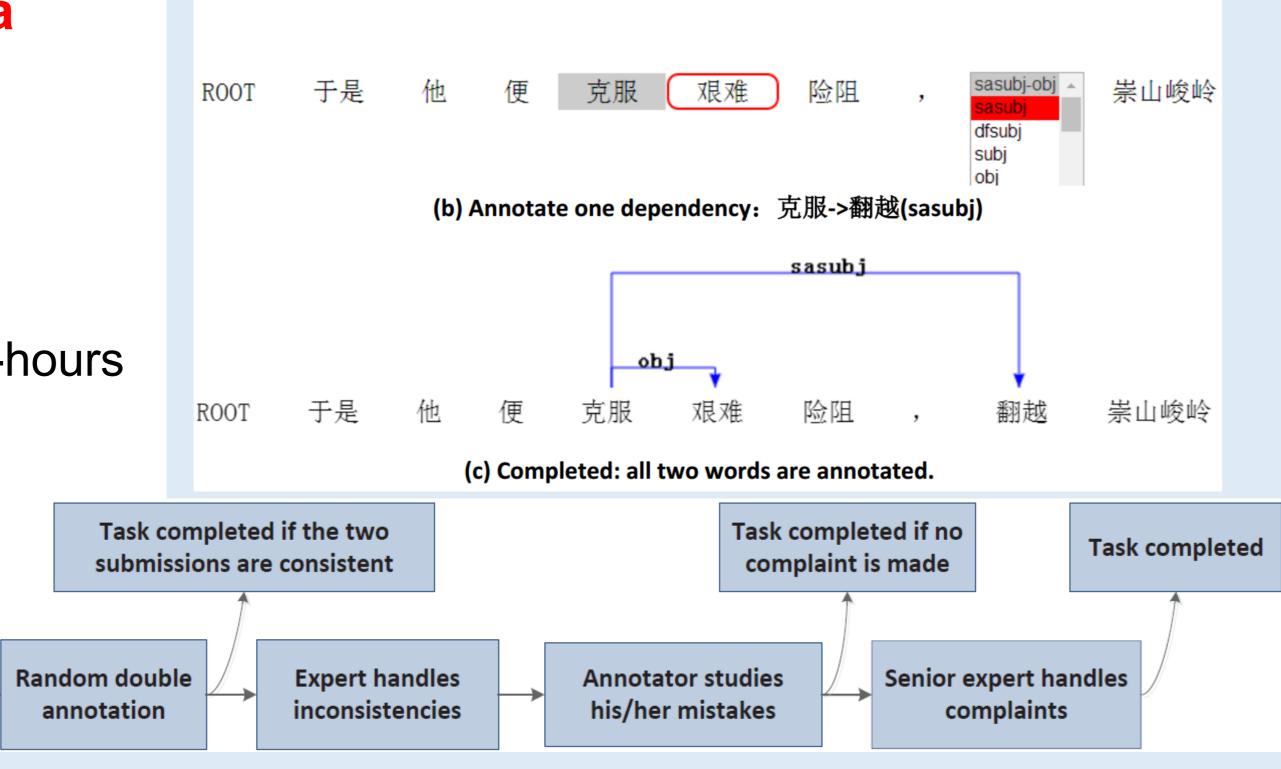
• Guideline (62 pages)

http://hlt.suda.edu.cn/index.php/SUCDT

Annotation platform

http://101.132.166.249/anno-sys

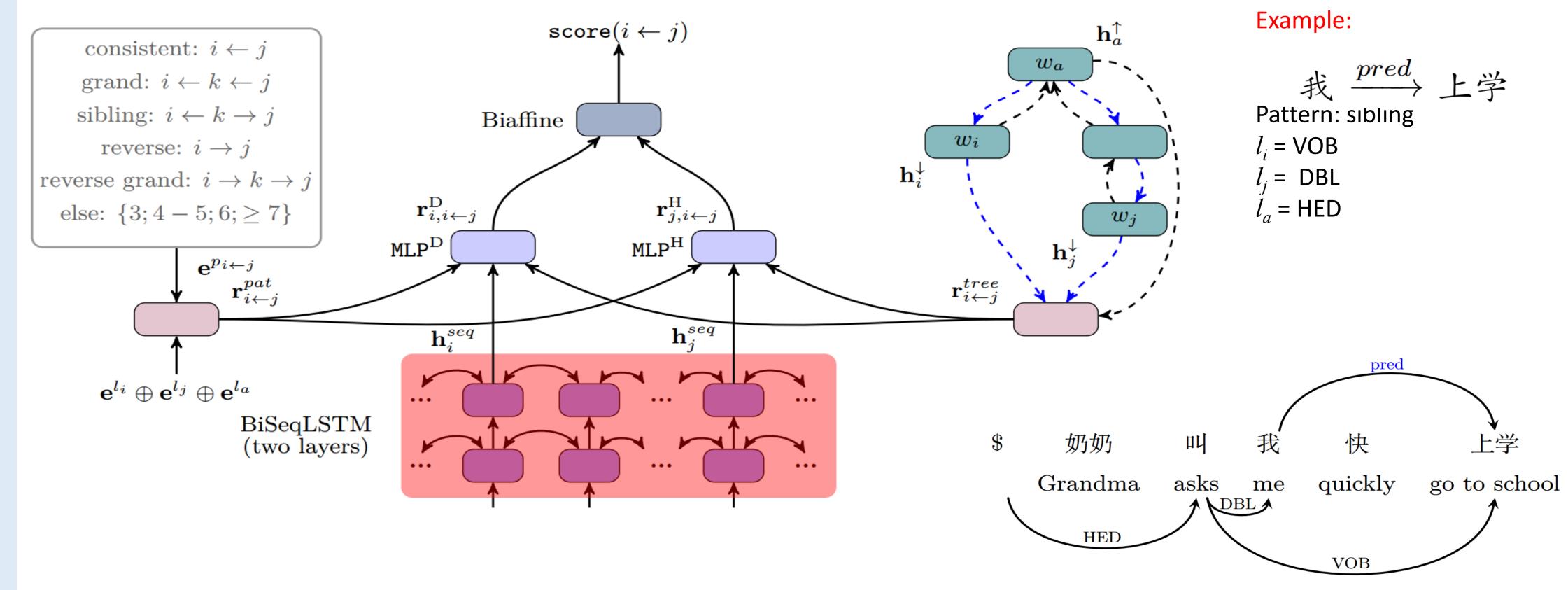
- Annotators: 15 part-time students
- Cost (~10K sentences): ~500 person-hours
- Overall Accuracy: 78.6%
- Inter-annotator consistency
- Dependency-wise: 76.5%
- Sentence-wise: 43.7%



(a) Before annotation: two words to be annotated

Approaches: pattern embedding & treeLSTM

• The key is how to make full use of the source-side tree as guidance.



Experiments

Data

	#Sent	#Tok (HIT)	#Tok (our)
train	7,768	119,707	36,348
dev	998	14,863	4,839
test	1,995	29,975	9,679
train-HIT	52,450	980,791	36,348

Parsing results on the test data

	Training data	UAS	LAS
Single	train	75.99	70.95
Single (hetero)	train-HIT	76.20	68.43
Multi-task	train & train-HIT	79.29	74.51
Single (large)	converted train-HIT	80.45	75.83

 Treebank conversion leads to better parsing accuracy than MTL.

Conversion results on the test data

	Training data	UAS	LAS
Multi-task	train & train-HIT	79.29	74.51
Pattern	train	86.66	82.03
TreeLSTM	train	86.69	$\bf 82.09$
Combined	train	86.66	81.82

- Treebank conversion is better than MTL due to the explicit use of the source-side tree.
- The two approaches are similarly effective.

Results on the fully annotated test data (372 sent)

Task	Training data	UAS LAS
Conversion	train	93.42 90.49
Parsing (baseline)	train	89.66 86.41
Parsing (ours)	converted train-HIT	91.16 88.07

• The real parsing and conversion accuracies are very high (over 90%).