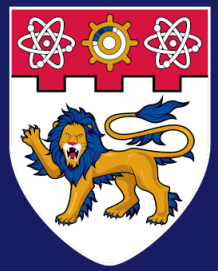


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ZHONG and Mal-rule Enhanced Treebanks

by

Luis Morgado da Costa

Contents

1. Background
2. ERG Mal-rule Enhanced Treebank / Maxent Model
3. ZHONG Coverage Updates
4. ZHONG Mal-rule Enhanced Treebank / Maxent Model
5. The Road Ahead

Background

- English and Mandarin Chinese Mal-rules
- NTU Corpus of Learner English (NTUCLE)
  - data collected from NTU engineering students assignments
- NTU Corpus of Learner Mandarin (NTUCLM)
  - data collected from NTU 1st year students
- LCC Automated Student Writing Support
  - supporting NTU students with their academic writing
  - blended-learning experiment showed positive results

ERG Mal-rule Enhanced Treebank

ERG Treebank

Data Sets		Treebankers					
ID	**Size**	**A**	**B**	**C**	**D**	**E**	**Overlap**
0	500	500	500	500	500	500	500
1	200	200	200				200
2	200			200	200		200
3	200	200					
4	200		200				
5	200					200	
6	200				200		
7	200	200					
8	200		200				
9	200					200	
10	200				200		
11	200	200				200	200
12	200		200		200		200
13	200					200	
14	200		200				
15	200					200	
16	200				200		
17	200					200	
18	200		200				
19	200					200	
20	200				200		
21	200				200	200	200
22	200		200			200	200
Total	4900	1300	1900	700	1900	2300	1700

ERG Treebank

ID	Size	Overlap					LA	UA
0	500	A	B	C	D	E	0.681	0.747
1	200	A	B				0.738	0.778
2	200			C	D		0.690	0.730
11	200	A				E	0.773	0.812
12	200		B		D		0.773	0.820
21	200				D	E	0.775	0.816
22	200		B			E	0.761	0.807
	1700						0.731	0.780

Labeled / Unlabeled Agreement based on PARSEVAL

- 76.3% of the sentences got a suitable tree
- Fully Open, a lot more of untagged data;
- Trained maxent model using ACE-Tools (GP=3);

Maxent: Grammaticality Judgments

	Parsed w/o errors	Parsed w/ errors	No parse
edERG (orig.)	0.589	0.315	0.096
edERG (new)	0.703	0.201	0.096
ERG	0.920	0.001	0.079
ERG → edERG (orig.)	0.921	0.037	0.042
ERG → edERG (new)	0.921	0.037	0.042

Parsing results of top/best parses for the test set (n=1000)

- problems defined by `_rbst` or the `_mal` suffixes (except `w_hasnoninitcap_dlr_rbst`);
- 349 sentences diagnosed as problematic by any system;

Maxent: Grammaticality Judgments

	Correctly Problematic	Incorrectly Problematic	Ignored Problematic	Correctly Ignored
edERG (orig.)	**0.413**	**0.490**	0.020	0.077
edERG (new)	**0.361**	**0.215**	0.072	**0.352**
ERG	0.003	0.000	0.430	0.567
ERG → edERG (orig.)	0.095	0.011	**0.338**	0.556
ERG → edERG (new)	0.095	0.011	**0.338**	0.556

Measured against 349 sentences diagnosed as problematic by any system;

	Precision	Recall	F1
edERG (orig.)	0.457	0.954	0.618
edERG (new)	0.627	0.834	**0.716**
ERG	1.000	0.007	0.013
ERG → edERG (orig.)	0.892	0.219	0.351
ERG → edERG (new)	0.892	0.219	0.351

Measured against parses w/rgd to predicted errors

Maxent: Diagnoses

	Correct Diagnosis	Incorrect Diagnosis	Missed Diagnosis
edERG (orig.)	0.530	**0.424**	0.046
edERG (new)	**0.642**	**0.192**	0.166
ERG	0.007	0.000	0.993
ERG → edERG (orig.)	0.146	0.073	0.781
ERG → edERG (new)	**0.185**	0.033	0.781

Measured against ungrammatical portion of the sentences (n=151)

	Precision	Recall	F1
edERG (orig.)	0.556	0.920	0.693
edERG (new)	**0.770**	0.795	**0.782**
ERG	1.000	0.007	0.013
ERG → edERG (orig.)	0.667	0.157	0.254
ERG → edERG (new)	**0.848**	**0.192**	0.313

Measured against parses w/rgd to error diagnoses

ZHONG Coverage Updates

ZHONG Development and Evaluation Sets

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

	Set Name	Set Size	Avg. Sent. Length
Development Data	cmnedu	798	7.74
	tufs	1531	6.46
	hsksc_01	175	5.71
	hsksc_02	200	7.92
	hsksc_03	81	9.42
	hsksc_04	200	10.51
	hsksc_05	200	11.89
	hsksc_06	157	13.48
	ntuclm_dev	2013	7.00
Evaluation Data	hsksc_07	200	16.77
	hsksc_08	200	19.84
	hsksc_09	30	22.23
	hsksc_10	200	21.71
	hsksc_11	200	23.12
	hsksc_12	67	22.97
	tatoeba_01	10000	8.47
	tatoeba_02	10000	7.95
	tatoeba_03	10000	7.94
	tatoeba_04	10000	7.44
	tatoeba_05	7216	7.23
	ntuclm_test	287	7.28

Extending Zhong

Use data collected by linguistic surveys:

- Constraining spurious ambiguity
- Tokenization issues
- Adding mal-rules
- Increase lexical and syntactic coverage
  - Separable verbs (说话 *to speak*, 见面 *to meet*)
  - Locative Expressions
  - Numbers/Classifiers/Numeral Predicates
  - New aspect hierarchy (interactions with negation)
  - etc.

Parsing Coverage - Development Data

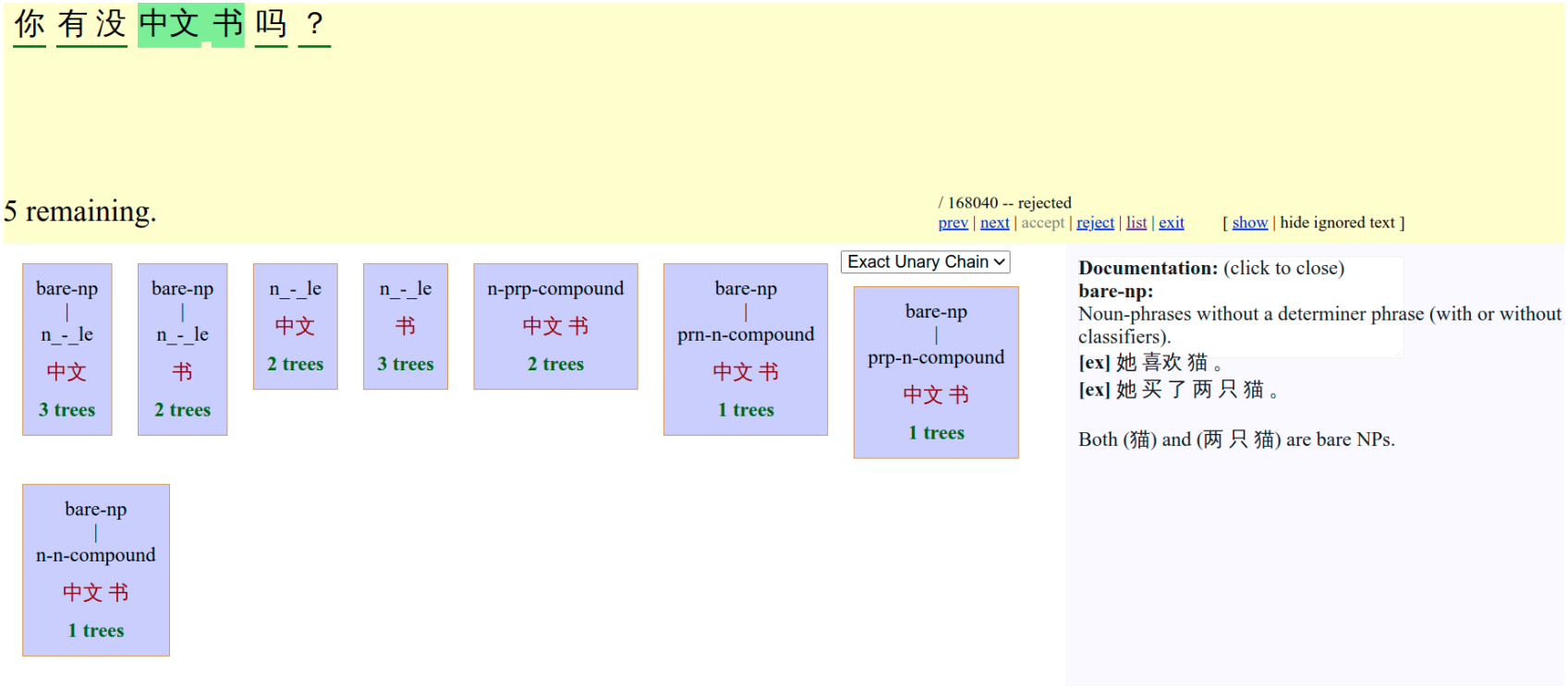
	ZHONG (v1.0)			ZHONG (v2.0)					
Set Name	% Parsed	Avg. Sent. Amb.	Avg. Sent. Len.	% Parsed	Δ	Avg. Sent. Amb.	Δ	Avg. Sent. Len.	Δ
cmnedu	91.9	261	6.6	94.9	+3.0	411.7	+150.4	6.7	+0.1
tufs	60.6	235	3.6	80.8	+20.2	222.7	-12.4	4.9	+1.3
hsksc_01	82.9	83	4.6	97.1	+14.3	50.7	-32.2	5.5	+0.9
hsksc_02	66.0	344	4.9	86.0	+20.0	319.2	-24.7	6.5	+1.7
hsksc_03	64.2	1332	5.5	76.5	+12.4	894.6	-437.4	6.6	+1.0
hsksc_04	56.5	843	5.0	70.5	+14.0	808.4	-34.5	6.5	+1.5
hsksc_05	48.0	1283	4.2	63.0	+15.0	1318.2	+35.2	6.4	+2.1
hsksc_06	46.5	4089	4.9	58.6	+12.1	2516.5	-1572.2	6.2	+1.3
ntuclm_dev	65.4	43	4.4	**92.4**	**+27.0**	244.01	+201.3	6.4	+2.0

Parsing Coverage - Evaluation Data

	ZHONG (v1.0)			ZHONG (v2.0)					
Set Name	% Parsed	Avg. Sent. Amb.	Avg. Sent. Len.	% Parsed	Δ	Avg. Sent. Amb.	Δ	Avg. Sent. Len.	Δ
hsksc_07	30.0	2422	2.6	35.0	+5.0	2151	-271	3.3	+0.7
hsksc_08	24.0	4281	2.9	32.0	+8.0	3273	-1009	4.1	+1.2
hsksc_09	26.7	1980	3.4	33.3	+6.7	5859	+3878	4.8	+1.4
hsksc_10	31.5	6070	4.1	33.0	+1.5	4573	-1497	4.2	+0.2
hsksc_11	20.0	5240	2.7	23.5	+3.5	5308	+68	3.5	+0.8
hsksc_12	16.4	3528	2.8	19.4	+3.0	2587	-940	2.9	+0.1
tatoeba_01	31.7	279	2.3	44.6	+12.9	2890	+10	3.3	+1.0
tatoeba_02	33.3	246	2.3	47.0	+13.7	220	-26	3.3	+1.0
tatoeba_03	31.8	185	2.3	46.2	+14.4	196	+11	3.3	+1.1
tatoeba_04	31.9	179	2.1	48.5	+16.6	144	-35	3.2	+1.2
tatoeba_05	29.2	145	1.9	50.5	+21.4	110	-34	3.4	+1.5
ntuclm_test	58.9	17	4.1	**91.6**	**+32.8**	211	+193	6.7	+2.5

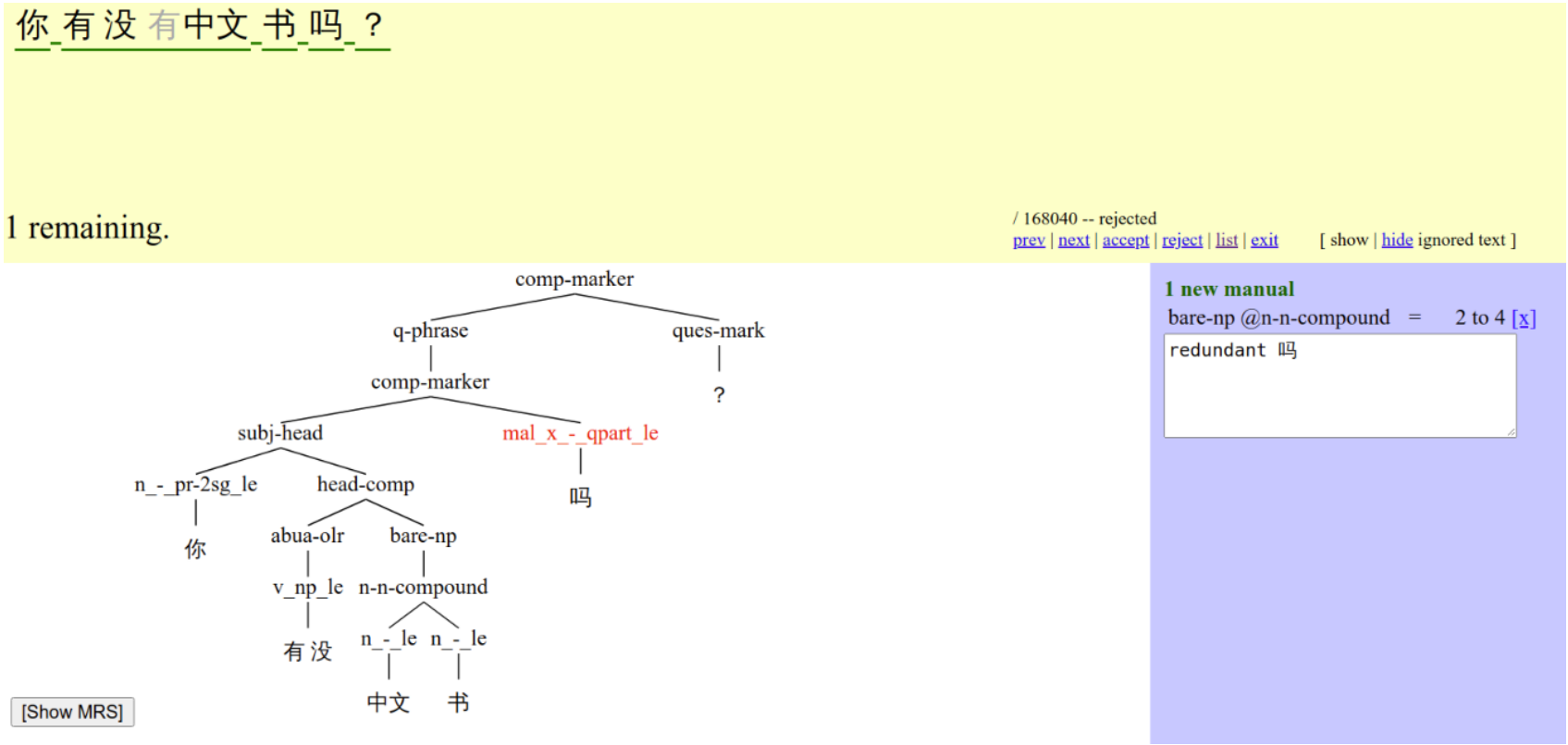
ZHONG Mal-rule Enhanced Treebank

Enhanced FFTB



In-tool display of LTDB-style grammar documentation through a JSON file;

Enhanced FFTB



Show the full name of rules, highlight mal-rules;

ZHONG Treebank

ID	Size	Overlap					LA	UA
tufs_cmn_01	200	**A**	**B**				0.870	0.897
tufs_cmn_02	200			**C**	**D**	**E**	0.795	0.840
tufs_cmn_03	200	**A**	**B**			**E**	0.880	0.905
tufs_cmn_04	200			**C**	**D**		0.817	0.848
tufs_cmn_05	200			**C**	**D**	**E**	0.839	0.900
tufs_cmn_06	200	**A**	**B**				0.877	0.928
tufs_cmn_07	200			**C**	**D**		0.839	0.867
tufs_cmn_08	137	**A**	**B**			**E**	0.874	0.892
cmnedu_01	200	**A**	**B**			**E**	0.824	0.873
cmnedu_02	200			**C**	**D**		0.779	0.820
cmnedu_03	200	**A**	**B**			**E**	0.851	0.884
cmnedu_04	198			**C**	**D**		0.801	0.834
hsksc_01	175	**A**	**B**			**E**	0.832	0.882
hsksc_02	200			**C**	**D**		0.775	0.832
hsksc_03	81	**A**	**B**			**E**	0.691	0.736
hsksc_04	200			**C**	**D**		0.791	0.826
hsksc_05	200	**A**	**B**			**E**	0.788	0.813
hsksc_06	157			**C**	**D**		0.767	0.794
ntuclm_test_01	200	**A**	**B**			**E**	0.794	0.817
ntuclm_test_02	87			**C**	**D**		0.624	0.642
ntuclm_train_01	200			**C**			-	-
ntuclm_train_02	200	**A**	**B**			**E**	0.874	0.900
ntuclm_train_03	200			**C**			-	-
ntuclm_train_04	200	**A**	**B**			**E**	0.871	0.897
ntuclm_train_05	200			**C**			-	-
ntuclm_train_06	200	**A**	**B**			**E**	0.884	0.912
ntuclm_train_07	200			**C**	**D**		0.808	0.832
ntuclm_train_08	200	**A**	**B**			**E**	0.859	0.885
ntuclm_train_09	200			**C**	**D**		0.533	0.543
ntuclm_train_10	213	**A**	**B**			**E**	0.721	0.733
Total	**5648**	**2806**	**2806**	**2842**	**2242**	**2806**	**0.808**	**0.893**

ZHONG Treebank

ID	Size	Treebanked	Parsed	Treebanked/ Parsed
tufs_cmn	1531	0.685	0.808	0.848
cmn_edu	798	0.896	0.949	0.945
hsksc_01	175	0.897	0.971	0.924
hsksc_02	200	0.665	0.860	0.773
hsksc_03	81	0.568	0.765	0.742
hsksc_04	200	0.580	0.705	0.823
hsksc_05	200	0.425	0.630	0.675
hsksc_06	157	0.357	0.586	0.609
ntuclm_train	2013	0.824	0.924	0.891
ntuclm_test	287	0.805	0.916	0.878
Total	5642	0.753	0.865	0.870

- Unfortunately, not all sets can be openly released;
- Trained maxent model using ACE-Tools (GP=3);

Maxent: Top Parse Against Treebank

System	Labeled Precision	Unlabeled Precision	Labeled Recall	Unlabeled Recall	Labeled F1	Unlabeled F1
ZHONG v1.0 (-)	0.439	0.591	0.422	0.566	0.431	0.578
ZHONG v1.0 (orig.)	0.470	0.601	0.449	0.573	0.459	0.587
ZHONG v2.0 (-)	0.780	0.934	0.812	0.984	0.796	0.959
ZHONG v2.0 (orig.)	0.920	0.977	0.927	0.986	0.924	0.981
ZHONG v2.0 (new)	**0.972**	**0.991**	**0.971**	**0.990**	**0.972**	**0.990**

Measured against NTUCLM Evaluation Data (n=287); Includes both grammatical and ungrammatical data;

Maxent: Grammaticality Judgments

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

	Correctly Problematic	Incorrectly Problematic	Ignored Problematic	Correctly Ignored
ZHONG v1.0 (-)	0.000	0.000	0.317	0.683
ZHONG v1.0 (orig.)	0.000	0.000	0.317	0.683
ZHONG v2.0 (-)	0.153	0.220	0.164	0.463
ZHONG v2.0 (orig.)	0.101	0.059	0.216	0.624
ZHONG v2.0 (new)	**0.129**	**0.017**	0.188	0.666

Measured against NTUCLM Evaluation Data

	Precision	Recall	F1
ZHONG v1.0 (-)	1.000	0.000	0.000
ZHONG v1.0 (orig.)	1.000	0.000	0.000
ZHONG v2.0 (-)	0.411	0.484	0.444
ZHONG v2.0 (orig.)	0.630	0.319	0.423
ZHONG v2.0 (new)	**0.881**	**0.407**	**0.556**

Measured against parses w/rgd to predicted errors

Maxent: Diagnoses

	Correct Diagnosis	Incorrect Diagnosis	Missed Diagnosis
ZHONG v1.0 (-)	0	0	1.000
ZHONG v1.0 (orig.)	0	0	1.000
ZHONG v2.0 (-)	0.143	0.341	0.516
ZHONG v2.0 (orig.)	0.275	0.044	0.681
ZHONG v2.0 (new)	**0.363**	**0.044**	0.593

Measured against ungrammatical portion of the NTUCLM Evaluation Data

	Precision	Recall	F1
ZHONG v1.0 (-)	1.000	0.000	0.000
ZHONG v1.0 (orig.)	1.000	0.000	0.000
ZHONG v2.0 (-)	0.295	0.302	0.299
ZHONG v2.0 (orig.)	0.862	0.357	0.505
ZHONG v2.0 (new)	**0.892**	**0.471**	**0.617**

Measured against parses w/rgd to error diagnoses

The Road Ahead

The Road Ahead

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- I need to finish writing my thesis

  'Any minute now...'

- Marie Skłodowska-Curie IF, European Council
  - Funded for 2 years;
  - Palacký University Olomouc, Czechia;
  - Work with Joanna Sio;
  - Mandarin NP Structure and Common Errors;

Thank you!