VLSP 2019 Shared Task: Dependency Parsing

NGUYEN Thi Minh Huyen

Hanoi - 2019

Outline

Outline

Dependency parsing

- Determining syntactic dependencies between words in a sentence: relationship between a predicate and its arguments, or a word and its modifiers
- Many applications: information extraction, co-reference resolution, question-answering, semantic parsing, ...
- Two shared tasks for Multilingual Parsing from Raw Text to Universal Dependencies
  - CoNLL shared task 2017[1]
  - CoNLL shared task 2018[2]

[1] http://universaldependencies.org/conll17/
[2] https://universaldependencies.org/conll18/

Vietnamese dependency parsing

- In 2008, N.L. Minh et al: MST parser on a corpus consisting of 450 sentences
- In 2013, N.T. Luong et al.: MaltParser on a Vietnamese dependency treebank
- In 2014, N. Q. Dat et al.: a new conversion method to automatically transform a constituent-based VietTreebank into dependency trees
- In 2017, N. K. Hieu: built BKTreebank, a dependency treebank for Vietnamese

Vietnamese dependency parsing

- In 2017, *a Vietnamese dependency treebank of 3,000 sentences is included for the CoNLL shared-task "Multilingual Parsing from Raw Text to Universal Dependencies"*: 48 dependency labels for Vietnamese based on Stanford dependency labels set
  - Small
  - Contains several errors

VLSP 2019: Vietnamese dependency parsing shared task

Outline

Data Preparation

- Training dataset
  - Viettreebank
  - The Little Prince
- Test dataset: Public test and Private test
  - Viettreebank
  - The Little Prince
  - Reviews on social media

Data Preparation

- Word segmentation
  - Annotation revision of 3 datasets
- Part of speech tagging
  - Update new POS labels
  - Map to the UPOS label set
- Definition of the new set of dependency relations
- Data annotation

Data Preparation

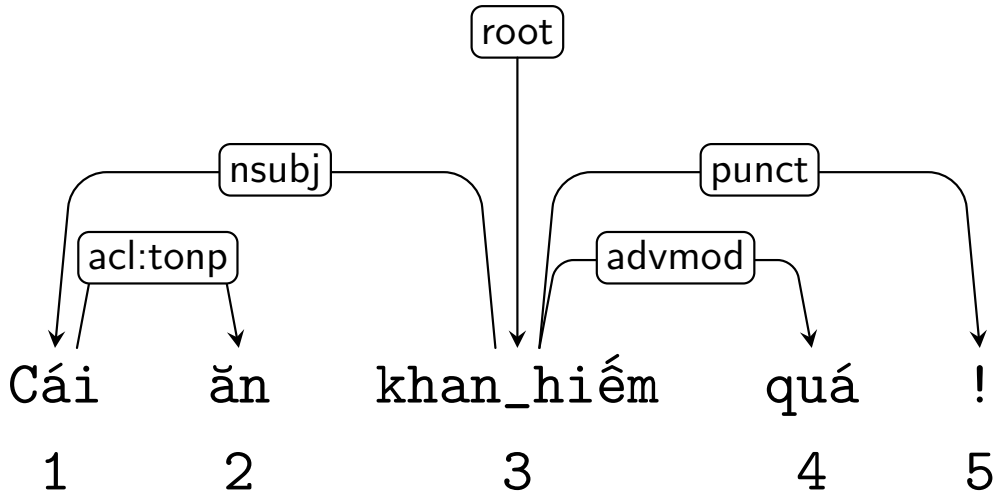
Dependency relation definition

- Based on universal dependency relations (UD V2)
  - 38 main relations
  - 47 subtypes

Data Preparation

Dependency relation definition

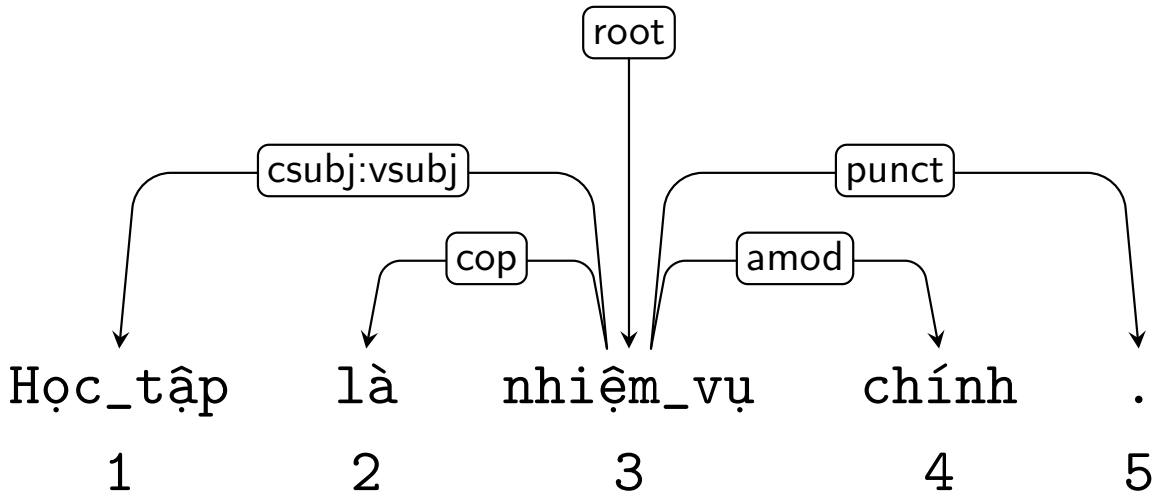
- Some new dependency labels specific to Vietnamese language:
  - *acl:tonp*: Verb nominalization using a classifier such as “cái”, “việc”, “sự”, ...



Data Preparation

Dependency relation definition

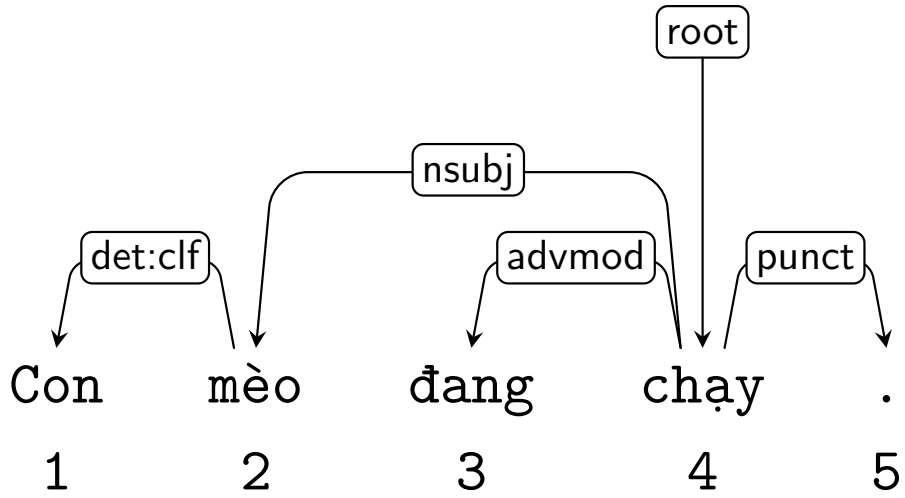
- Some new dependency labels specific to Vietnamese language:
  - *csubj:vsubj*



Data Preparation

Dependency relation definition

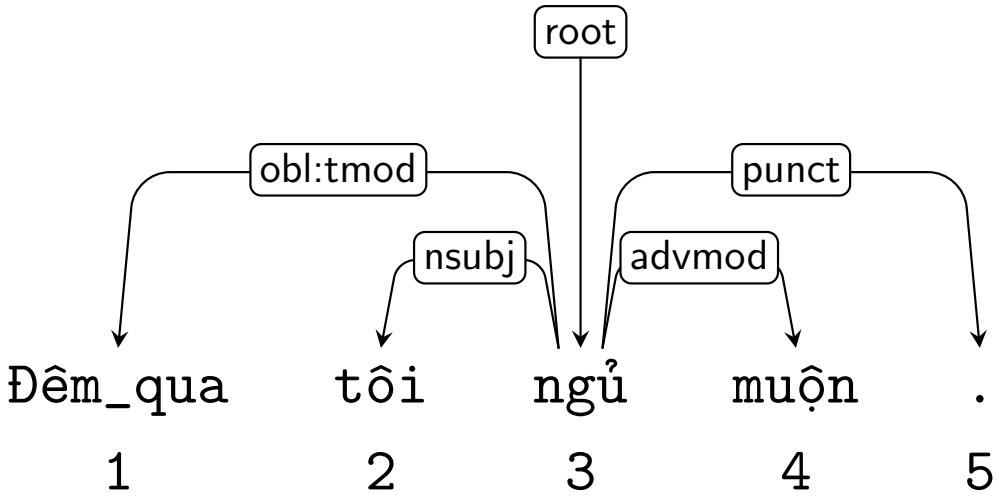
- Some new dependency labels specific to Vietnamese language:
  - *det:clf* :



Data Preparation

Dependency relation definition

- Some new dependency labels specific to Vietnamese language:
  - *obl:tmod*



Data Preparation

Data annotation

- 5 annotators: 2 months
- Tool

Outline

Data format

- The dependency-annotated data must be encoded in CoNLL-U format

A sentence in training dataset

1	Tôi	tôi	PROPN	Pro	_	3	nsubj	_	_
2	đã	đã	ADV	Adv	_	3	advmod	_	_
3	sống	sống	VERB	V	_	0	root	_	_
4	nhiều	nhiều	ADJ	Adj	_	3	advmod:adj	_	_
5	với	với	SCONJ	C	_	7	case	_	_
6	những	những	DET	Det	_	7	det	_	_
7	người lớn	người lớn	N	N	_	3	obl:with	_	_
8	.	.	PUNCT	PUNCT	_	3	punct	_	_

Data statistics

Number of sentences and average words per sentence

No.	Dataset	Sentences	AvgWS
1	Training Dataset1	2920	14.58
2	Training Dataset2	935	11.29
3	Public Test	100	13.61
4	Private Test 1 - MXH	100	12.14
5	Private Test 2 - VTB	400	24.82
6	Private Test 3 - HTB	100	11.00

Data statistics

Number of labels in each dataset

No.	Dataset	Total Label	Main Label
1	Training Dataset1	81	34
2	Training Dataset2	75	35
3	Public Test	64	33
4	Private Test 1 - MXH	56	31
5	Private Test 2 - VTB	74	34
6	Private Test 3 - HTB	53	33

Evaluation metrics

- UAS: the percentage of words that are assigned correct syntactic head
- LAS: the percentage of words that are assigned both the correct syntactic head and the correct dependency label

$$P = \frac{correctRelations}{systemNodes}$$

$$R = \frac{correctRelations}{goldNodes}$$

$$LAS = \frac{2 * P * R}{(P + R)}$$

Outline

Results

- 15 registered teams
- 4 teams submitted results for public test
- 3 teams submitted results for final test

Methods

- DP1 used Stanford graph-based neural dependency parser
  - Made a few modifications to adapt for Vietnamese dependency parsing
  - Investigated three models using different hyperparameters for the optimizer in training
- DP2 proposed a joint model for POS tagging and dependency parsing
  - Consists of a BiLSTM-CNN-CRF-based POS tagger and a Deep Biaffine Attention based dependency parser
  - A combined objective function is used to jointly train both models
- DP3 developed a simple ensemble model for Vietnamese dependency parsing task
  - Used two probability layers of the deep biaffine attention parser method with two different pre-trained word embeddings

Result

Public test result

Public test

Team	LAS
DP1	64.22
DP2	60.84
DP3	68.33
DP4	70.76

Result

Private test result

UAS metric

UAS	HTB	MXH	VTB
DP1-Model1	77.00	63.26	65.15
DP1-Model2	78.45	66.23	68.23
DP1-Model3	81.55	67.63	69.93
DP2-Model1	81.36	67.22	68.07
DP2-Model2	81.73	65.65	68.21
DP2-Model3	81.55	66.47	68.93
DP3-Model1	81.73	67.46	**72.95**
DP3-Model2	**85.91**	**67.79**	72.85

Result

Private test result

LAS metric

LAS	HTB	MXH	VTB
DP1-Model1	65.73	51.81	52.36
DP1-Model2	68.55	54.53	54.99
DP1-Model3	72.64	56.67	57.46
DP2-Model1	73.55	55.52	55.69
DP2-Model2	72.91	54.12	55.95
DP2-Model3	72.55	54.86	56.61
DP3-Model1	72.73	**57.08**	**60.19**
DP3-Model2	**77.09**	56.75	60.08

Result

Private test result

UAS and LAS on all 3 datasets

Team	UAS	LAS
DP1-Model1	66.03	53.5
DP1-Model2	68.95	56.16
DP1-Model3	70.75	58.75
DP2-Model1	69.18	57.28
DP2-Model2	69.17	57.29
DP2-Model3	69.82	57.87
DP3-Model1	73.19	61.01
DP3-Model2	**73.53**	**61.28**

Outline

Awards

- First rank
  Duc-Vu Nguyen, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen
  UIT-VNUHCM
- Second rank
  Thi Thuy Lien Nguyen and Quang Nhat Minh Pham
  Aimesoft JSC
- Third rank
  Xuan-Dung Doan
  VTCC