Overcoming Early Saturation on Low-Resource Languages in Multilingual Dependency Parsing

Jiannan Mao[†], Chenchen Ding[‡], Hour Kaing[‡],
Hideki Tanaka[‡], Masao Utiyama[‡], Tadahiro Matsumoto[†]

[†]Gifu University, Gifu, Japan
[‡]National Institute of Information and Communications Technology, Kyoto, Japan

Outline

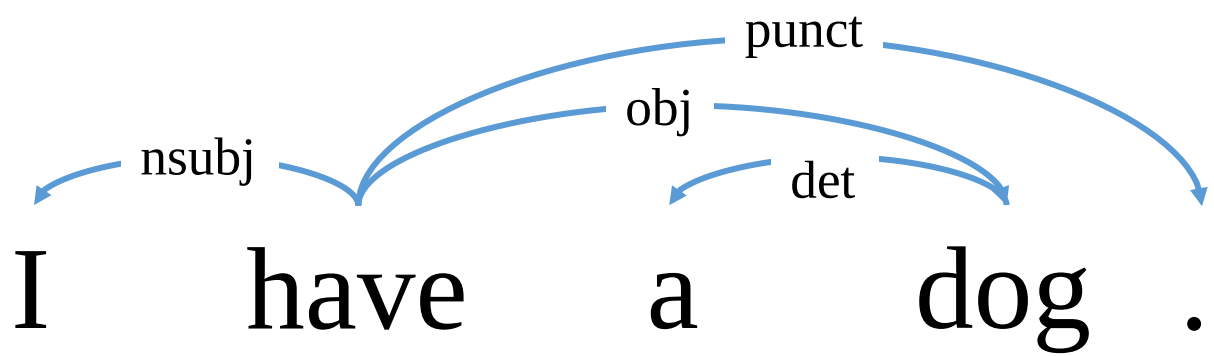
- Overview
- Background
- Investigation
- Experiments
- Conclusion and Future Work

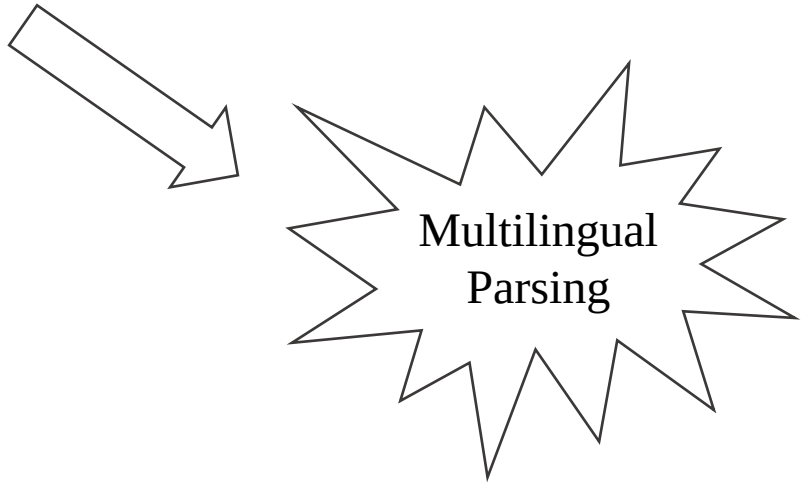
Dependency Parsing

- Dependency parsing:
  1. linguistic analysis technique
  2. uncover grammatical relationships(words in a sentence).

- Large treebanks: efficiently.

- For low-resource languages: treebanks are small/unavailable.





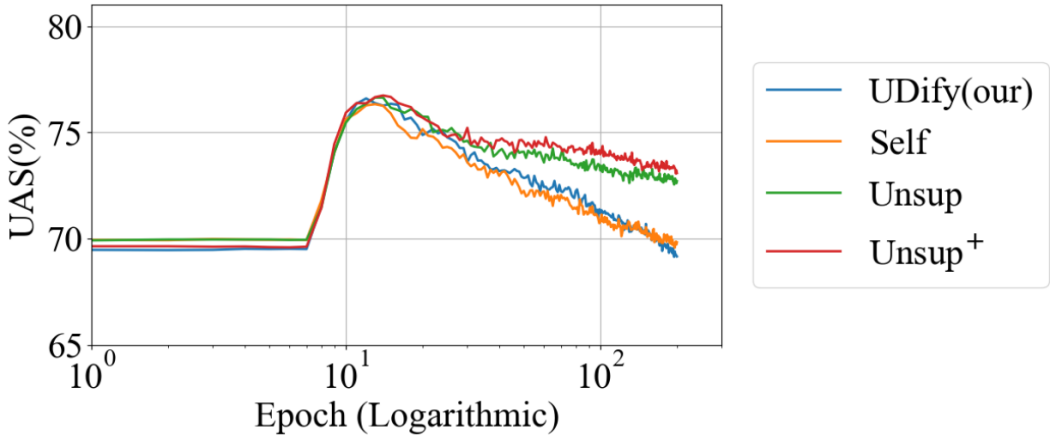
Multilingual Parsing and UDify

- A single model: parsing different languages.
- The lack of data: cross-lingual information.

- A multilingual multi-task parser.
- Exhibits strong and consistent performance in all UD treebanks.
- Early saturation occurs in some low-resource languages.

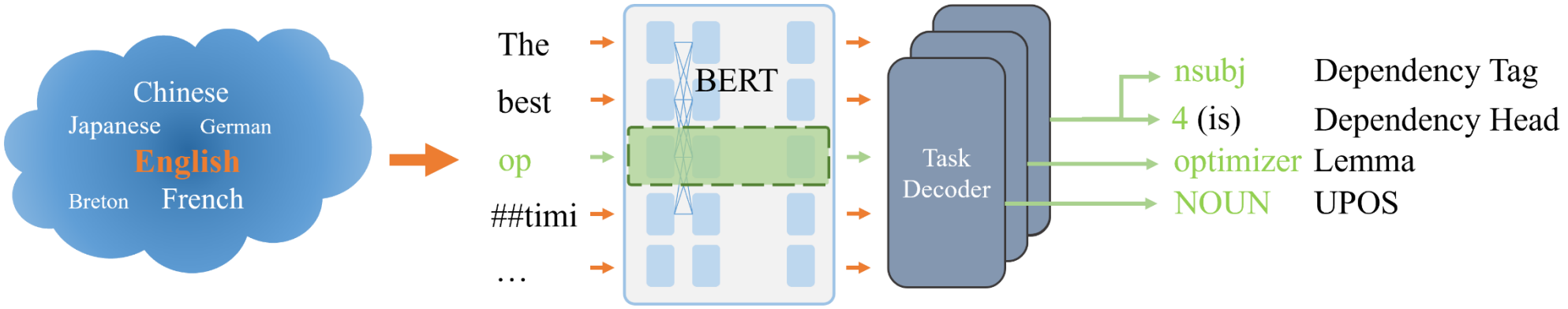
Contributions:

- on multiple low-resource languages using data augmentation methods.
- the unlabeled attachment score (UAS), stability: enhance
- Robustness of multilingualism processing is still retained.



UDify

- Lemmas, POS tags, and dependency structures.
- Finetuned on multilingual BERT.
- No language tag

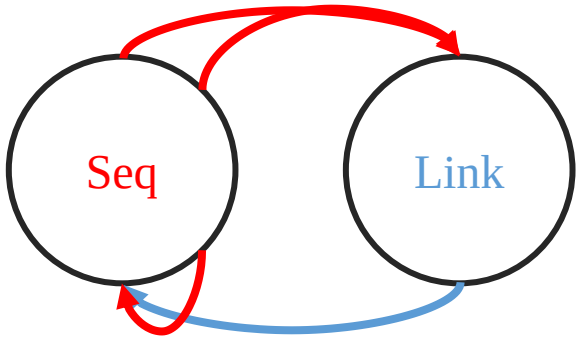


English raw sentence input example : The best optimeizer is grad student descent

Unsupervised Dependency Learning

- An unsupervised algorithm for dependency learning (Unsupervised-Dep).

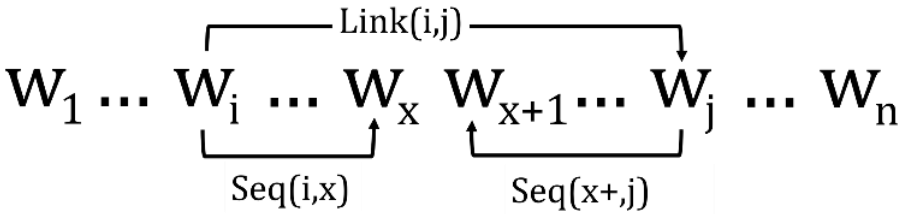
- Constructs the tree
    1. a dynamic programming method (CYK chart)
    2. the complete-link and complete-sequence

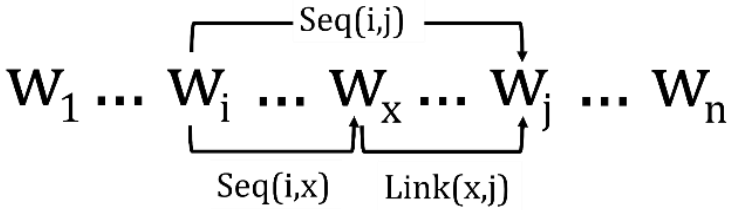


- Considering the time complexity of N-gram, focus on the **bi-gram**.

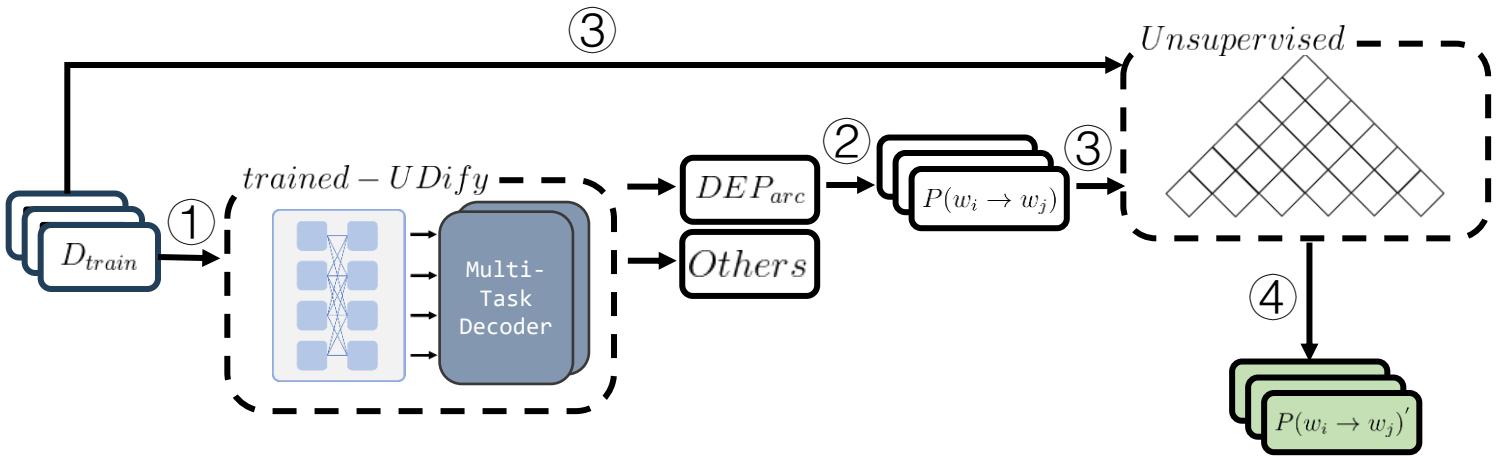
Unsupervised Dependency Learning: **bi-gram**

- Dependency relations define pair directions:
  $(w_i \rightarrow w_j)$, $(w_i \leftarrow w_j)$.
- Calculated using the Inside-Outside algorithm.
- Tree construction determined by Viterbi algorithm to ensure maximum probability.





UDify with Data Augmentation - Training



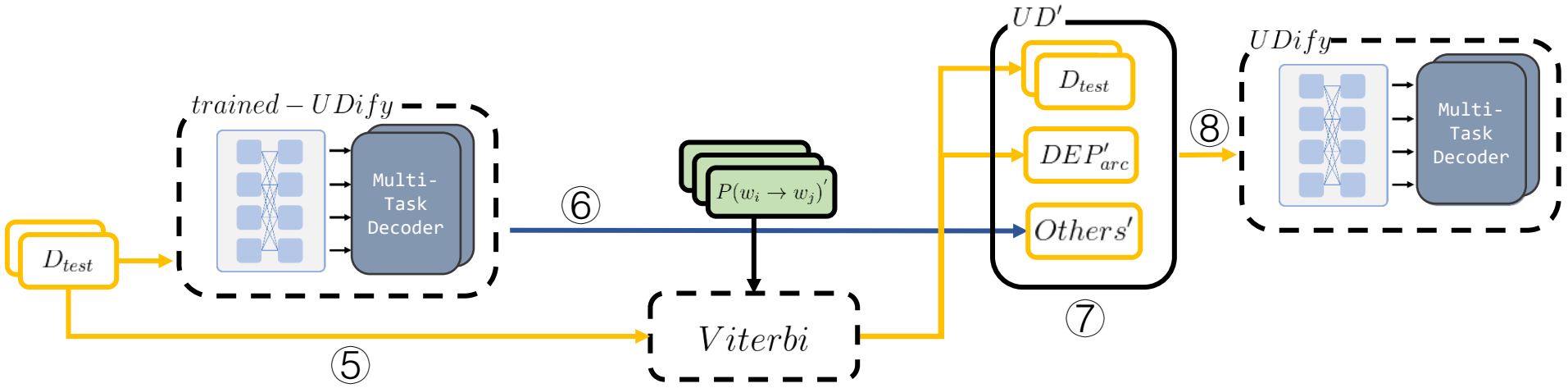
①Feed the D_{train}, into trained-UDify

②Statistical computations are performed on DEP_{arc}

③Unsupervised-Dep: $P(w_i \rightarrow w_j)$,$P(w_i \leftarrow w_j)$ and D_{train}

④Obtain the re-estimated probabilities $P(w_i \rightarrow w_j)'$ and $P(w_i \leftarrow w_j)'$

UDify with Data Augmentation - Generation



⑤Find the the optimal structure of D_{test} by $P(w_i \rightarrow w_j)'$, $P(w_i \leftarrow w_j)'$ and Viterbi

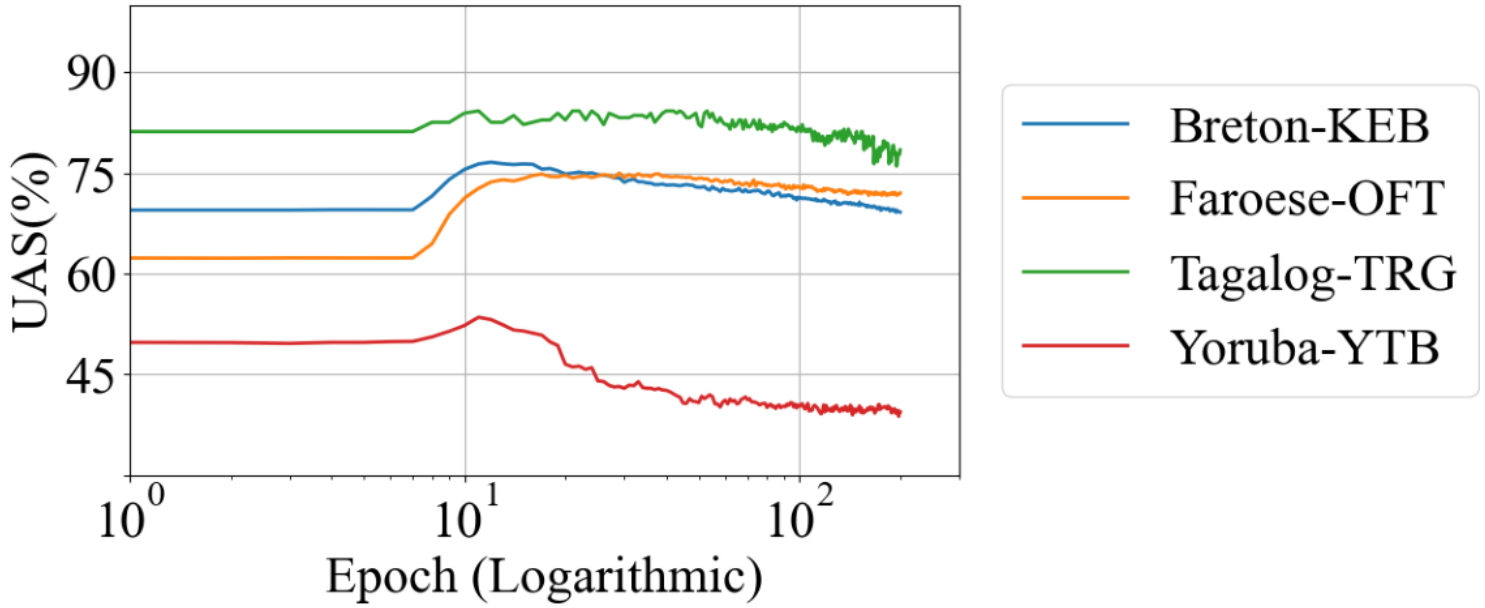
⑥Retain information other than DEP_{arc} of the parser result from UDify

⑦Construct the D_{test} in the UD treebank format

⑧Train a new UDify

On Few- and Zero-Shot Languages

- Early saturation in the accuracy of dependency parsing was observed.
- Unsupervised-Dep data augmentation across multiple low-resource languages remains underexplored.



Dataset and Setup

language(code)	#sent.(len.)	#train	#test
Armenian(hy)	2.4(8.2)	560	470
Belarusian(be)	2.0(9.0)	260	68
Hungarian(hu)	134.1(5.3)	910	449
Kazakh(kk)	1.7(8.2)	31	1,047
Lithuanian(lt)	236.7(5.6)	153	55
Marathi(mr)	1.5(10.0)	373	47
Tamil(ta)	13.7(7.7)	400	120
Breton(br)	18.2(9.5)	0	888
Faroese(fo)	1.3(8.1)	0	1,208
Tagalog(tl)	150.0(16.2)	0	55
Yoruba(yo)	9.7(8.1)	0	100

OPUS-mult: Raw data collected from various corpora.

- **UDify(our)**: Reproduced UDify model.
- **Unsup**: UD 2.3+OPUS-mult_{300} , generated by Unsupervised-Dep.
- **Self**: UD 2.3+ OPUS-mult_{300} , the parsing results from Baseline.

Dependency Task on the Low-Resource Languages

	hy	be	hu	kk	lt	mr	ta	br	fo	tl	yo	Few	Zero
UDify(org)	85.6	91.8	89.7	74.8	79.1	**79.4**	79.3	63.5	67.2	64.0	37.6	-	-
UDify(our)	86.1	92.1	89.8	76.0	79.4	74.3	80.8	69.2	72.0	78.4	39.4	84.0	67.1
Self	85.9	**92.5**	89.6	**76.2**	79.2	74.8	**81.2**	69.8	**72.5**	85.3	38.8	84.0	67.6
Unsup	**86.3**	92.4	**90.0**	**76.2**	**79.5**	74.0	80.5	**72.7**	71.9	**88.0**	**39.6**	**84.2**	**68.7**

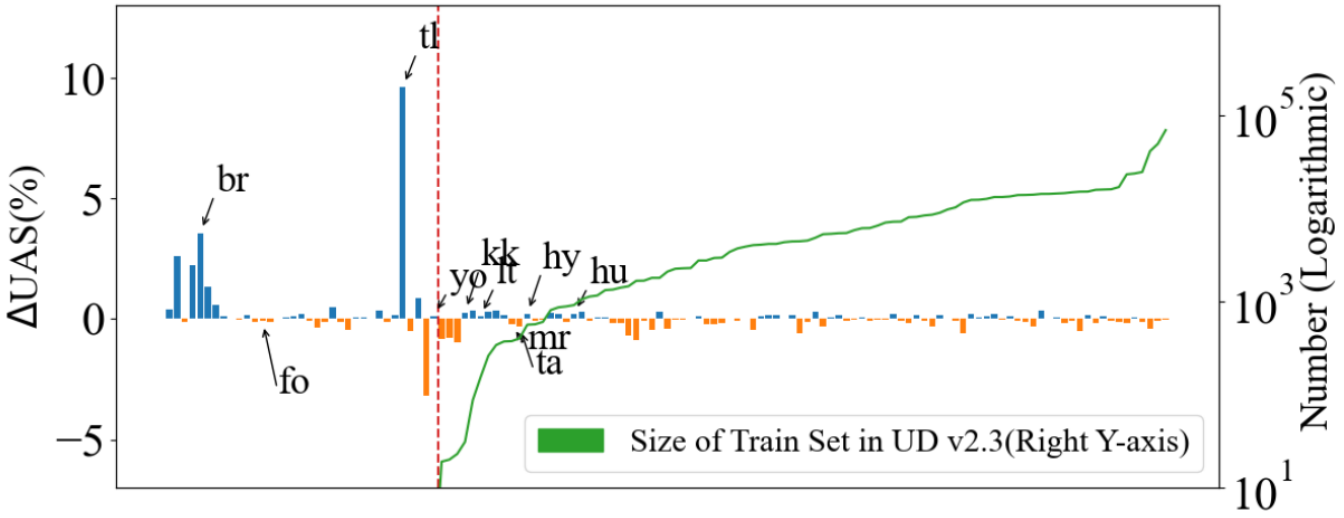
UAS for few- and zero-shot languages obtained using different methods.

few-shot languages: contain a little of training data in UD treebanks.
zero-shot languages: do not contain any training data in UD treebanks.

Other Tasks and Languages

	Zero-shot		Other	
	UAS	Rest	UAS	Rest
UDify(our)	67.1	55.6	77.5	82.5
Self	67.6	56.3	77.5	82.4
Unsup	**68.7**	**59.0**	77.5	82.5

UD scores on selected zero-shot and other languages obtained by different methods.



Difference in UAS on all test treebanks.
X-axis: sorted order of treebank training sets from smallest to largest.

- Employed data augmentation through unsupervised learning.
- Overcome early saturation in parsing accuracy among low-resource languages.
- Future Work:
  - exploring additional factors
  - using the latest UD treebanks

Thank you
for
Listening!