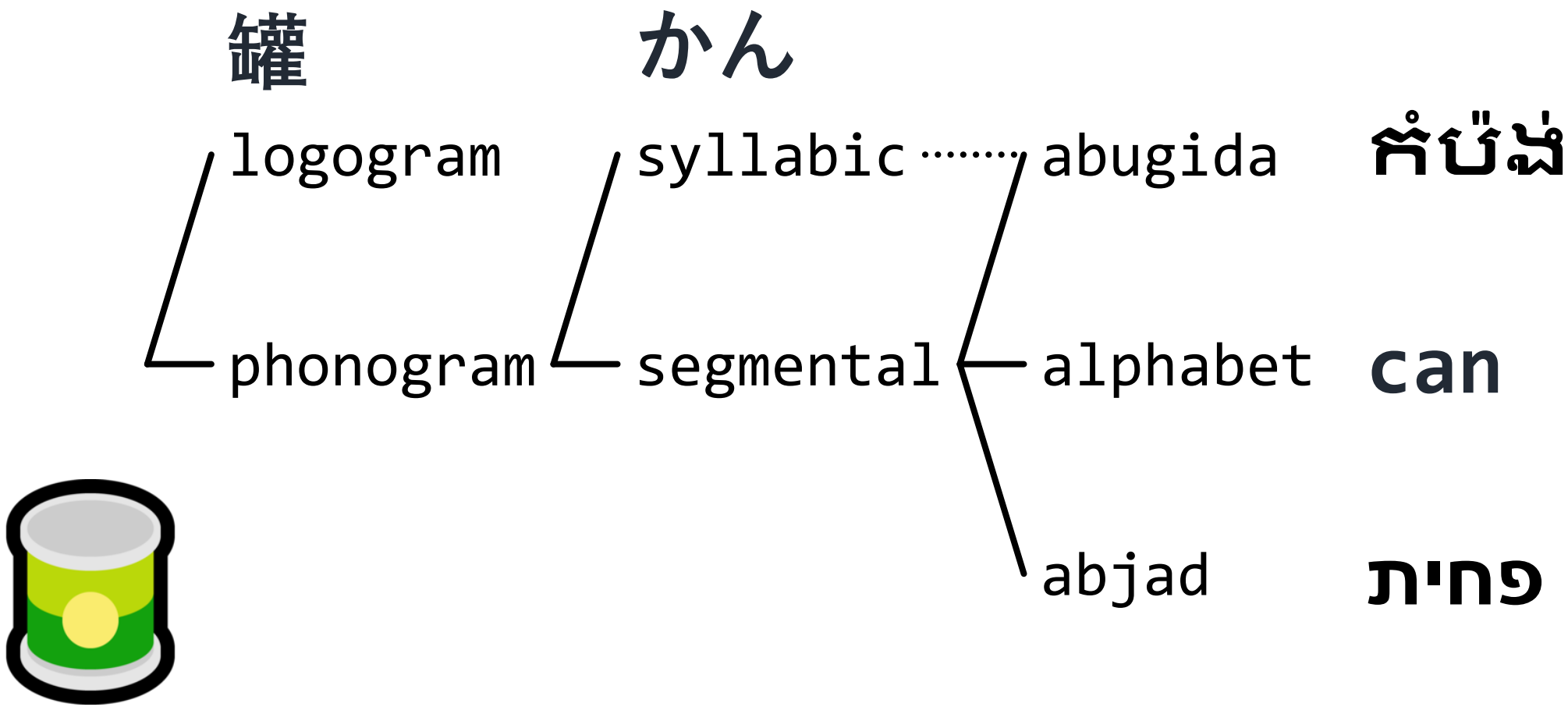
Simplified Abugidas

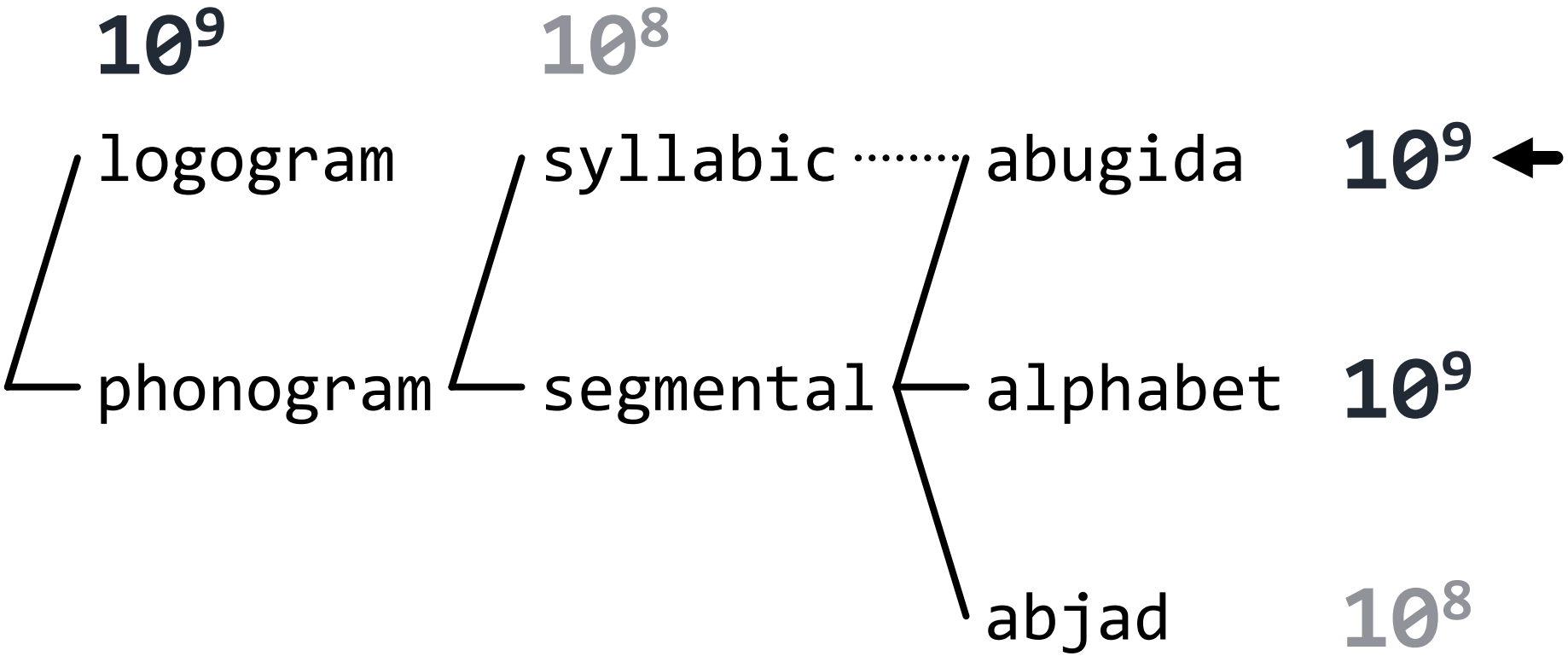
Chenchen Ding, Masao Utiyama, and Eiichiro Sumita

ASTREC, NICT, Japan

Writing Systems : Hierarchy



Writing Systems : Population



Writing Systems : How to Input

Logogram > Syllabic

Keyboard

↓

Input method

↓

10^2 – 10^4 symbols

> Abugida >

Alphabet & Abjad

Keyboard

↓

20 – 40 symbols

Writing Systems : How to Input

Logogram > Syllabic

Keyboard

↓

Input method

↓

10^2 – 10^4 symbols

> Abugida >

Keyboard

?

typically 50 – 70

Alphabet & Abjad

Keyboard

↓

20 – 40 symbols

Motivation of This Study

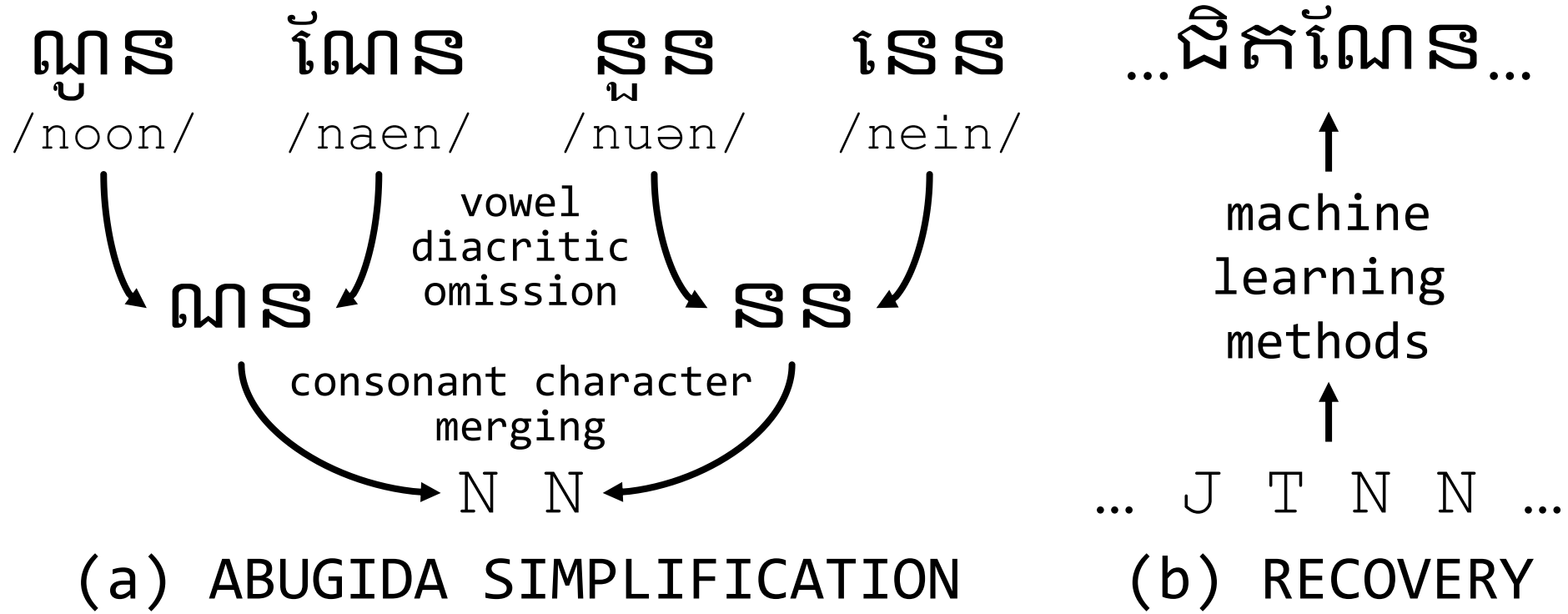
- Can abugidas be inputted more efficiently?
  → To insert a light layer of input method
  → To type less and to recover automatically

- Related work
  → Various approaches for Chinese and Japanese

  → To take advantage of redundancy in a writing system

Outline of This Study

- Khmer script as an example



Abugida Simplification

- Four Brahmic abugidas
  - Thai, Burmese (Myanmar), Khmer (Cambodian), and Lao
- Based on phonetics / conventional usages → reduced to 21 symbols

MERGED	GUTTURAL			PALATE				DENTAL				LABIAL				R-LIKE	S-LIKE	H-LIKE	ZERO-C.	LONG-A	PRE-V. DE-V.
	PLOSIVE		NAS.	PLOSIVE		NAS.	APP.	PLOSIVE		NAS.	APP.	PLOSIVE		NAS.	APP.						
	I	II		I	II			I	II			I	II								
MN	K	G	U	C	J	I	Y	T	D	N	L	P	B	M	W	R	S	H	Q	A	E
TH	กขฃ	คฅฆ	ง	จฉ	ชซฌ	ญ	ย	ฎฏฐดตถ	ฑฒทธ	ณน	ลฦฬ	บปผฝ	พฟภ	ม	ว	รฤ	ศษส	หฮ	อ	าๅ	เแโใไ
MY	ကခ	ဂဃ	င	စဆ	ဇဈ	ဉ ည	ယ ◌ျ	ဋဌတထ	ဍပဒဓ	ဏန	လဠ	ပဖ	ဗဘ	မ	ဝ ◌ွ	ရ ◌ြ	သဿ	ဟ ◌ှ	အ	◌ါ ◌ာ	◌္ ◌်
KM	កខ	គឃ	ង	ចឆ	ជឈ	ញ	យ	ដឋតថ	ឌឍទធ	ណន	លឡ	បផ	ពភ	ម	វ	រ	ឝឞស	ហ	អ	◌ា	◌្
LO	ກຂ	ຄ	ງ	ຈ	ຊ	ຍ	ຢ	ດຕຖ	ທ	ນ	ລ	ບປຜຝ	ພຟ	ມ	ວ	ຣ	ສ	ຫຮ	ອ	າ	ເແໂໃໄ

OMITTED		
	TH	ะ ◌ั ◌ำ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู ◌ฺ ◌็ ◌่ ◌้ ◌๊ ◌๋ ◌์ ◌ํ ◌๎
	MY	◌ိ ◌ီ ◌ု ◌ူ ေ◌ ◌ဲ ◌ံ ◌့ ◌း
	KM	◌ិ ◌ី ◌ឹ ◌ឺ ◌ុ ◌ូ ◌ួ ◌ើ ◌ឿ ◌ៀ េ◌ ែ◌ ៃ◌ ោ◌ ៅ◌ ◌ំ ◌ះ ◌ៈ ◌៉ ◌៊ ◌់ ◌៌ ◌៍ ◌៎ ◌៏ ◌័ ◌៑
	LO	ະ ◌ັ ◌ຳ ◌ິ ◌ີ ◌ຶ ◌ື ◌ຸ ◌ູ ◌ົ ◌ຼ ຽ ◌່ ◌້ ◌໊ ◌໋ ◌໌ ◌ໍ

Abugida Simplification

- Khmer script as an example

$$\text{ជិតណែន} = \text{ជ} + \text{ិ} + \text{ត} + \text{ណ} + \text{ែ} + \text{ន} \rightarrow \text{J T N N}$$

$$\text{Leng.} = \frac{\text{len}(\text{"J T N N"})}{\text{len}(\text{"ជ ិ ត ណ ែ ន"})} = \frac{4}{6} = 66.7\%$$

	Thai	Burmese	Khmer	Lao
Leng.	76.0%	74.0%	77.6%	72.8%

- Around **one quarter** characters (**diacritics**) saved

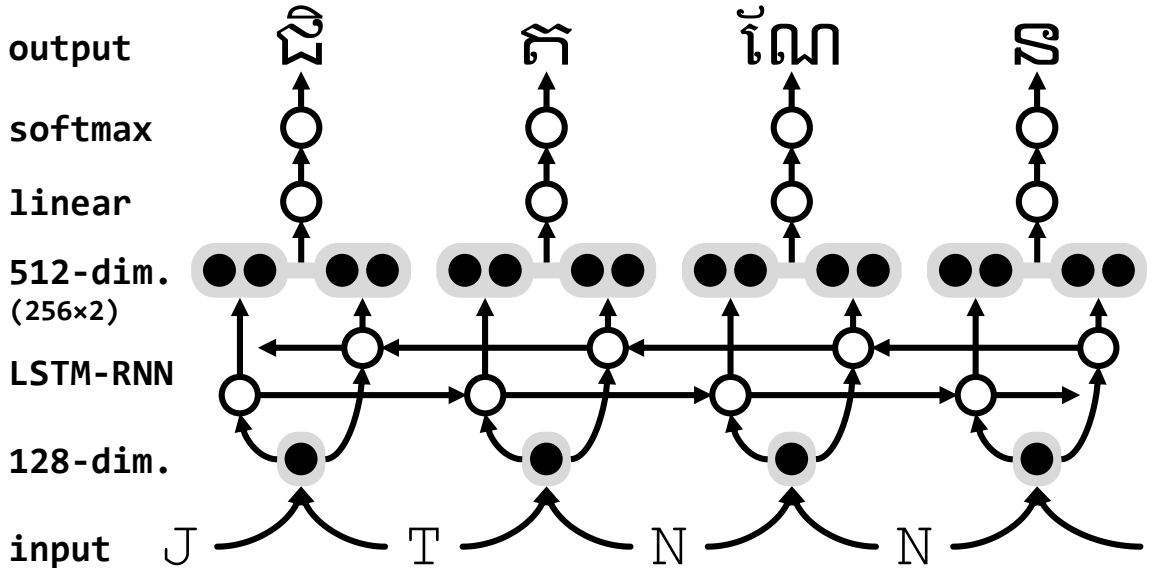
Recovery Methods

- To formulate as a sequential labeling task
  - However, list-wise search as in **conditional random fields** is costing
- To solve by point-wise classification
  - **Support vector machine (SVM)** as a baseline
  - **Recurrent neural network (RNN)** as a state-of-the-art method
- Setting for the SVM baseline
  - Linear kernel with N-gram features
  - Using **LIBLINEAR** library
    - Wrapped by the **KyTea** toolkit

RNN Structure and Settings

- Bi-gram of graphemes as input
- **Embedding** → **Bi-directional LSTM** → Linear transform → Softmax
- Original writing units as output

- Implemented by **DyNet**
- Trained by **Adam**
  - Initial learning rate 0.001
  - Controlled by a validation set
- Multi-model ensemble



Experimental Results

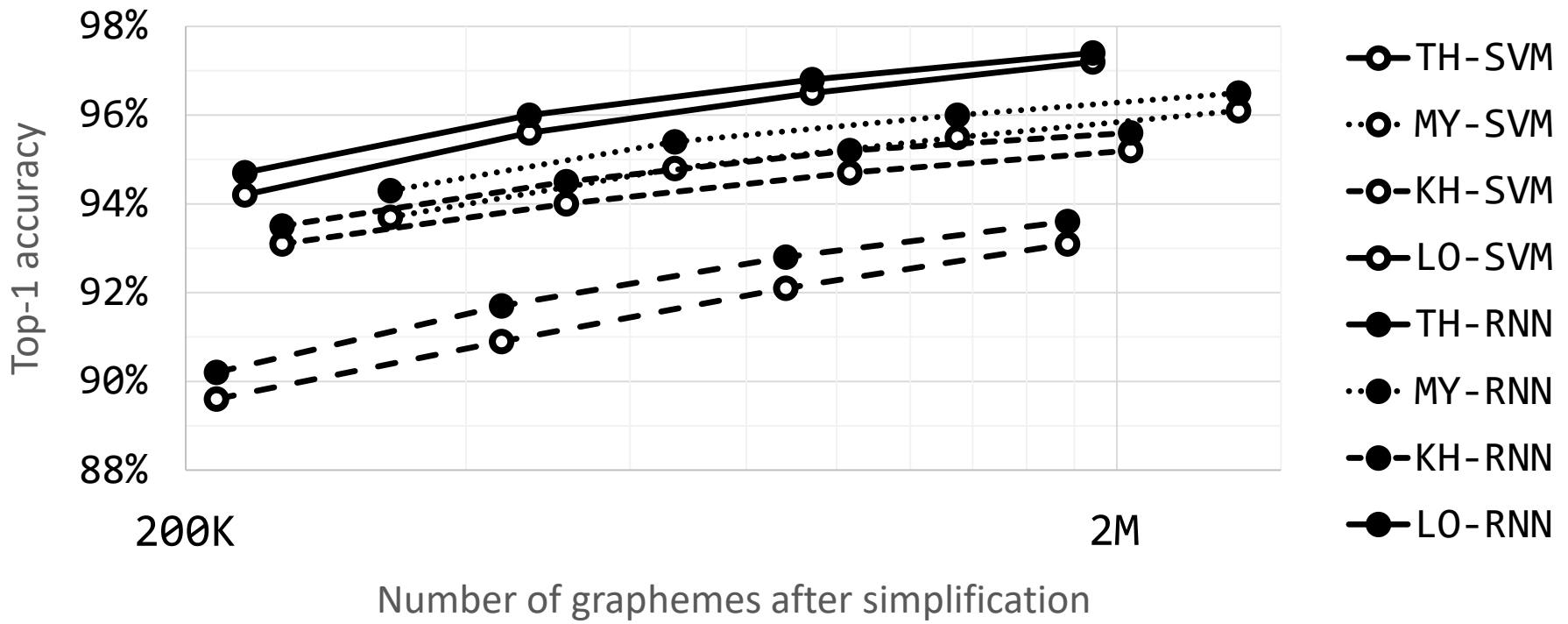
- Asian Lang. Treebank data
  - 20,000 sent., newswire
- **SVM**
  - Up to 5-gram for TH, KM, LO
  - 7-gram for Burmese
- **RNN**
  - **⊕M** : M-ensemble
    → **16**-ensemble is adequate
  - **@N** : Top-N results
    → Top-**4** is satisfactory

	Thai	Burmese	Khmer	Lao
SVM	97.2%	96.1%	95.2%	93.1%
$\text{RNN}^{@1}_{\oplus 4}$	97.2%	96.3%‡	95.5%‡	93.3%‡
$\text{RNN}^{@1}_{\oplus 8}$	97.3%†	96.4%‡	**95.6%**‡	**93.6%**‡
$\text{RNN}^{@1}_{\oplus 16}$	**97.4%**‡	**96.5%**‡	**95.6%**‡	**93.6%**‡
$\text{RNN}^{@2}_{\oplus 16}$	98.8%	98.4%	97.5%	96.6%
$\text{RNN}^{@4}_{\oplus 16}$	99.2%	98.8%	98.1%	97.7%
$\text{RNN}^{@8}_{\oplus 16}$	99.2%	98.9%	98.4%	97.9%

† : p < 0.01, ‡ : p < 0.001

→ Embedding + bi-LSTM > N-gram features

Experimental Results : Training Data Size



→ RNN outperforms SVM, regardless of the training data size

Manual Evaluation

- On Burmese and Khmer best results by RNN
- Conducted by native-speakers
- To classify errors into four-level
  - `0. acceptable, i.e., alternative spelling`
  - `1. clear and easy to identify the correct result`
  - `2. confusing but possible to identify the correct result`
  - `3. incomprehensible`

Level	0	1	2	3
Burmese	4.5%	51.0%	42.2%	2.2%
Khmer	22.5%	28.5%	16.3%	32.8%

Conclusion and Future Work

- Abugidas can be simplified largely and recovered with high accuracy
  - Four Brahmic abugidas are investigated
  - Simplified into a compact symbol set (around 20 graphemes)
  - Recovered satisfactorily by standard machine learning method

  → Experimentally show the feasibility to encode abugidas in a lossy manner
- Future work
  - Language specific investigation
  - To integrate dictionary
  - To develop practical input method for abugidas

Thanks for your kind attention