ICMLA 2019

Language-Agnostic Syllabification with Neural Sequence labeling

Jacob Krantz* Maxwell Dulin Paul De Palma

Department of Computer Science, Gonzaga University

Spokane, WA

jkrantz@jkrantz@zagmail.gonzaga.edu

GONZAGA UNIVERSITY | School of Engineering & Applied Science

**now at Oregon State University*

Preliminaries: The Syllable

- Phone: A unit of sound.
  - t in “ **t**ip ”
- Syllable: A single segment of uninterrupted phones
  - Highly debated among linguists
  - Not all words have a set syllable pattern; can be multiple

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2018)

CONSONANTS (PULMONIC)

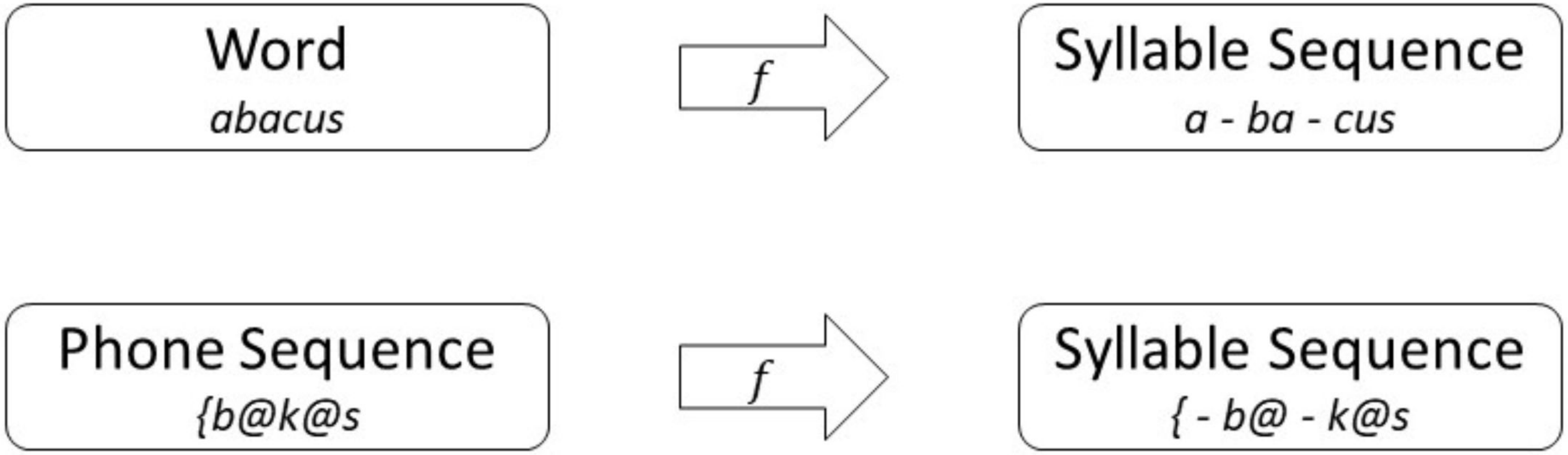
© 2018 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Chart: Wikimedia Commons

Preliminaries: Syllabification



Existing Automatic Approaches

Dictionary-Based

- Requires a lot of manual effort
- Cannot handle new words

Rule-Based

- Too many rules (exceptions, exceptions to exceptions, …)
- Not very accurate, rigid

NIST tsylb software package (*Fischer, 1996*)

- Implementation of Daniel Kahn’s 1979 MIT dissertation
- Around 3000 hand transcribed rules for English syllabification

Existing Automatic Approaches: Data-Driven

Strength:

- Learn the function *f* from examples
- No hand crafted linguistic knowledge needed outside of the dataset
- Given labeled data, can possibly learn *f* for any language

Challenge:

- Limited labeled training data.

Hidden Markov models (HMMs), support vector machines (SVMs), and conditional random fields (CRFs)
(Demberg et al, 2006) *(Bartlett et al, 2009)* *(Singh et al, 2016)*

Contributions

1. Developed a unique and general neural network architecture for data-driven syllabification that achieves or competes with state of the art language-specific models.

2.

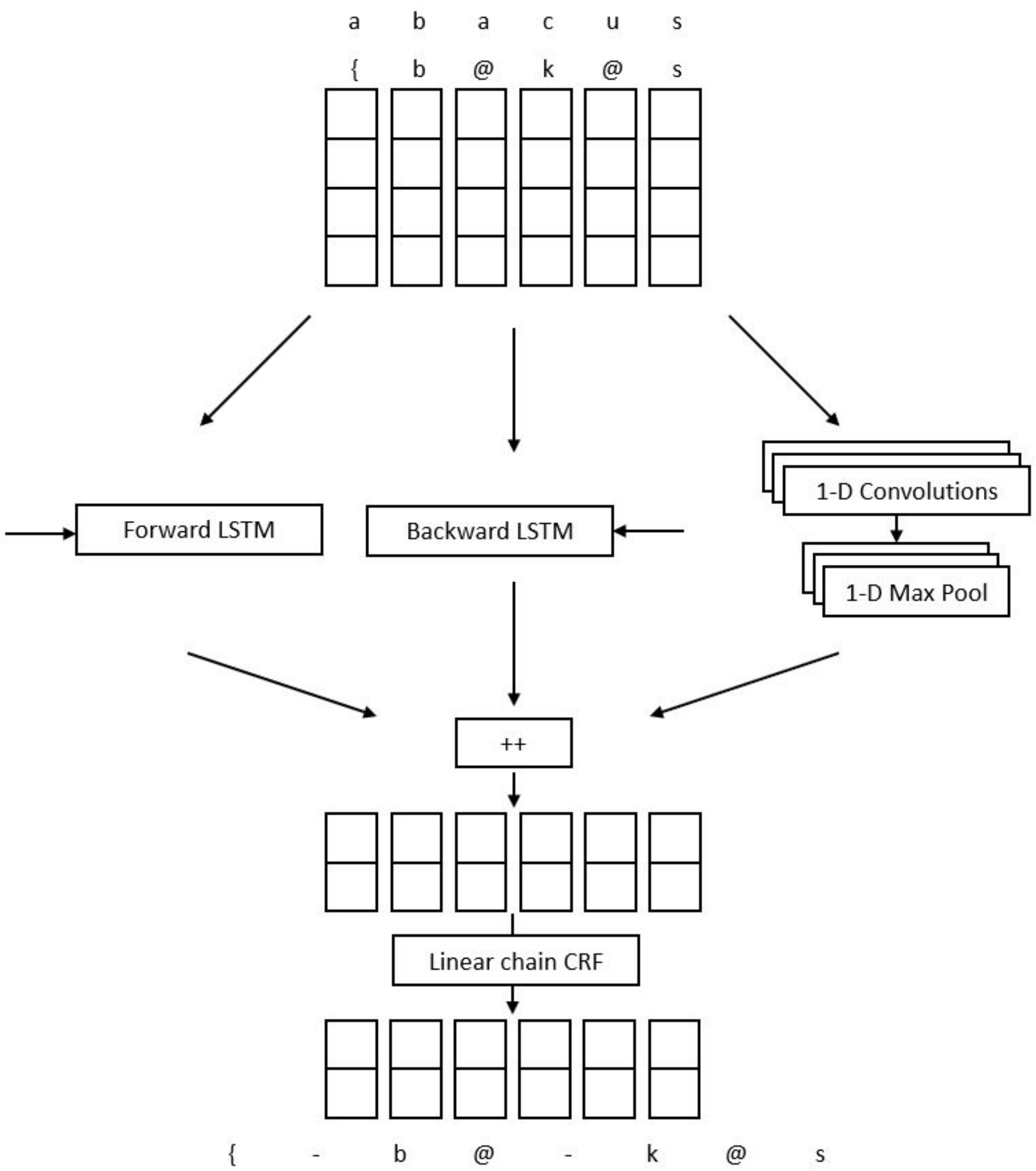
Method

Treat as a labeling task

Components:

- Phone Embeddings
- Bi-LSTM
- CNN
- Linear-Chain CRF

Network Architecture Diagram



Method: Prediction

$$\hat{y} = \arg\max_{y} p(y|o)$$

Option 1: Softmax

$$p(y|o) \approx \prod_{i=0}^{n-1} \max(s_i)$$

$$s_i = \frac{e^{o_i}}{\sum_{k=0}^{1} e^{o_{i_k}}}$$

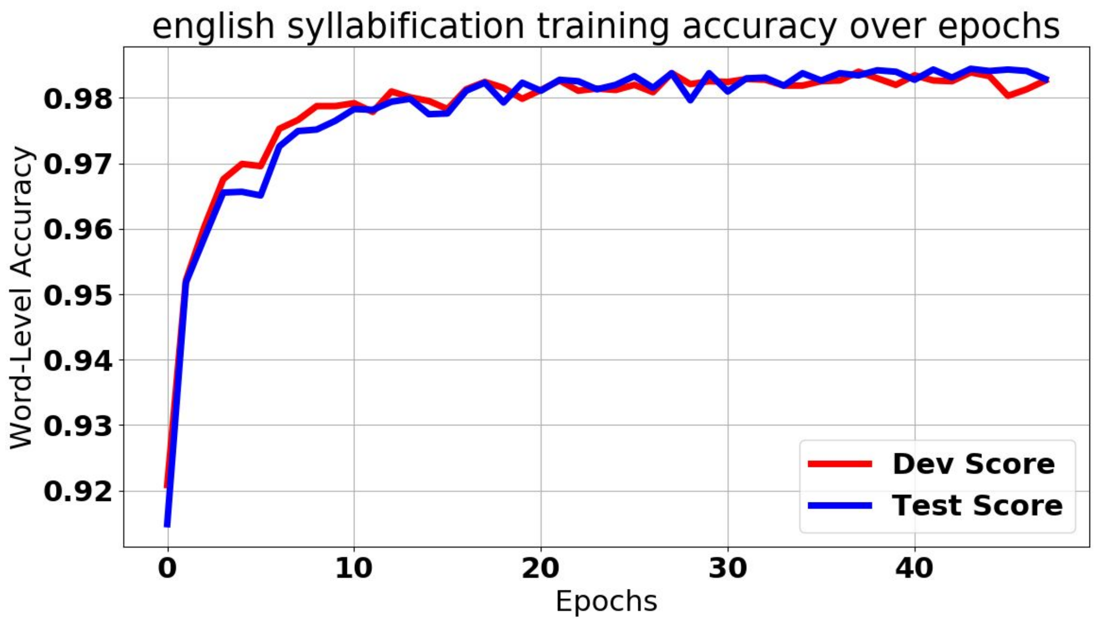
Option 2: Conditional Random Field

(Lafferty et al, 2001 & Huang et al, 2015)

$$p(y|o) \approx \frac{1}{Z(o)} \prod_{i=1}^{n-1} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_i, y_{i-1}, o_i)\right\}$$

Method: Training

- Minibatch by length
- Adam optimizer
- Early stopping



Training time on one GPU: 30-45 minutes (English dataset)

Contributions

1. Developed a unique and general neural network architecture for data-driven syllabification that achieves or competes with state of the art language-specific models.

2. Performed a more expansive evaluation across languages to better test language generalizability

Evaluation: Datasets & Languages

Indo-European: Romance

Language	Italian	French
Dataset	Festival (Taylor et al, 1998)	OpenLexique (New et al, 2004))
Words	440K	139K

Sino-Tibetan: Tibeto-Burman

Language	Manipuri
Dataset	IIT-Guwahati (Singh et al, 2016))
Words	17K

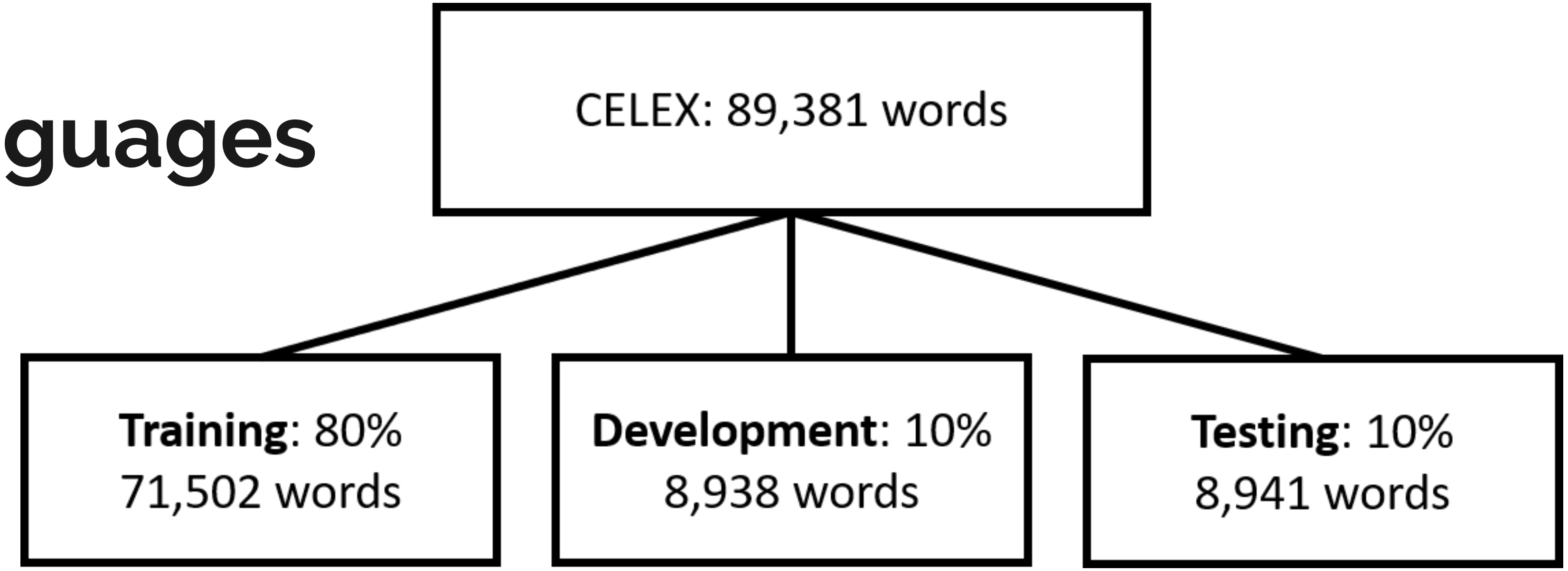
Indo-European: West Germanic

Language	English	Dutch
Dataset	CELEX (Baayen et al, 1995)	CELEX (Baayen et al, 1995)
Words	89K	328K

Language Isolate

Language	Basque
Dataset	E-Hitz (Perea et al, 2006)
Words	100K

Evaluation:
Datasets & Languages



Hyperparameters were tuned on the English CELEX dataset
Experiments were repeated 20 times

Results

Existing Syllabifiers

Dataset	Syllabifier	Method	%
English CELEX	*tsylb* [10]	Rule-based	93.72
English CELEX	HMM-GA [15]	Data-driven	92.54
English CELEX	Learned EBG [41]	Data-driven	97.78
English CELEX	SVM-HMM [13]	Data-driven	98.86
Dutch CELEX	SVM-HMM [13]	Data-driven	99.16
Festival	Liang hyphenation [40]	Data-driven	99.73
OpenLexique	Liang hyphenation [40]	Data-driven	99.21
IIT-Guwahat	Entropy CRF [36]	Hybrid	97.5
E-Hitz	Liang hyphenation [40]	Data-driven	99.68

Our Models

Model	English CELEX	Dutch CELEX	Festival	OpenLexique	IIT-Guwahat	E-Hitz
Base	$\mathbf{98.5 \pm 0.1}$	$\mathbf{99.47 \pm 0.04}$	$\mathbf{99.990 \pm 0.005}$	99.98 ± 0.01	94.9 ± 0.3	$\mathbf{99.83 \pm 0.07}$
Small	98.2 ± 0.2	99.39 ± 0.04	$\mathbf{99.990 \pm 0.004}$	99.987 ± 0.007	$\mathbf{95.4 \pm 0.3}$	99.68 ± 0.06
Base-Softmax	97.7 ± 0.2	99.24 ± 0.06	99.984 ± 0.003	$\mathbf{100.00 \pm 0.01}$	94.7 ± 0.3	99.71 ± 0.04

Highlighted Examples

Word	Generated	Target
misinterpretation	**mIs-In-t3-prI-t1-SH**	mIs-In-t3-prI-t1-SH
achieved	**@-Jivd**	@-Jivd
worrisome	**wV-rI-sF**	wV-rI-sF
public-address systems	pV-blI-k-@-d-rEs-sI-st@mz	pV-blIk-@-drEs-sI-st@mz

Phones in DISC format

- Successful with long words and various conjugations
- Struggled with hyphenation and spaces

Code available: https://github.com/jacobkrantz/lstm-syllabify

Conclusion

1. Developed a unique, general, and language-agnostic neural network architecture for data-driven syllabification
   a. Components: phone embeddings, BiLSTM, CNN, CRF
2. Performed a more expansive evaluation across languages to better test language generalizability

Performing both RNN and CNN processing over the same input can increase accuracy. (Ma & Hovy, 2016) showed this works for cnn over characters, rnn over words which is similar it different.

Going forward:

- Explore ways to harness the power of neural networks when faced with limited training data

References

- W. Fisher, “Tsylb syllabification package,” 1996, accessed 1 July, 2019 from: https://www.nist.gov/itl/iad/mig/tools.
- V. Demberg, “Letter-to-phoneme conversion for a german text-to-speech system,” Master’s thesis, University of Stuttgart, 2006.
- S. Bartlett, G. Kondrak, and C. Cherry, “On the syllabification of phonemes,” in Proceedings of NAACL-HLT: 2009. ACL, 2009, pp. 308–316.
- L. G. Singh, L. Laitonjam, and S. R. Singh, “Automatic syllabification for manipuri language,” in Proceedings of COLING 2016: Technical Papers, 2016, pp. 349–357.
- J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proceedings of the 18th ICML, 2001, pp. 282–289.
- Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” arXiv preprint arXiv:1508.01991, 2015.

Thank you!

- Funding provided by the McDonald Work Award, established generously by Robert and Claire McDonald
- AWS Cloud Credits for Research Program