Speech Communication 000 (2016) 1–10

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Indonesian syllabification using a pseudo nearest neighbour rule and phonotactic knowledge

Suyanto[a,b,*], Sri Hartati[a], Agus Harjoko[a], Dirk Van Compernolle[c]

[a] Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia
[b] School of Computing, Telkom University, Jl. Telekomunikasi Terusan Buah Batu, Bandung, West Java 40257, Indonesia
[c] Departement Elektrotechniek-ESAT, KU Leuven, Kasteelpark Arenberg 10, Leuven 3001, Belgium

ARTICLE INFO

Article history:
Received 7 January 2016
Revised 28 October 2016
Accepted 31 October 2016
Available online xxx

Keywords:
Indonesian syllabification
Four-feature phoneme encoding
Phonotactic knowledge
Pseudo nearest neighbour rule

ABSTRACT

This paper discusses phonemic syllabification using a pseudo nearest neighbour rule (PNNR) and phonotactic knowledge for Indonesian language. The proposed data-driven model uses a four-feature phoneme encoding and a phonotactic-based pre-syllabification. Evaluating on 50 k words dataset using 5-fold cross-validation shows that the proposed encoding significantly reduces the average syllable error rate (SER) by 13.90% relatively to the commonly used orthogonal binary encoding and the pre-syllabification also reduces the average SER up to 17.17% relatively to the PNNR without pre-syllabification. Five-fold cross-validating proves that the proposed PNNR-based syllabification is stable by producing an average SER of 0.64%. Most errors come from derivatives with the prefixes 'ber', 'per', and 'ter' as well as from compound words. This result is also significantly lower than a Look-Up-based syllabification that gives an average SER of 2.60%.

1. Introduction

There are two approaches to automatic syllabification: rule-based and data-driven (statistic-based). The rule-based approach, using the sonority sequencing principle, legality principle, and maximal onset principle, produces ambiguous syllabifications that are only valid for some cases (Rogova et al., 2013). This approach gives high accuracy for languages with few simple and consistent syllabification rules, such as Sinhala Weerasinghe et al. (2005) and Spanish Hernández-Figueroa et al. (2013). But, it performs poorly for a slightly more complex syllabic language, such as Malay Hafiz et al. (2011). Hence, many data-driven methods have been developed, such as the IGTree learning algorithm (Daelemans et al., 1997), weighted finite-state transducers (Kiraz et al., 1998), combination of treebank and bracketed corpora training (Müller, 2001), neural network (Daelemans and van den Bosch, 1992; Kristensen, 2000; Tian, 2004), probabilistic context-free grammars (Müller, 2006), joint n-gram models (Schmid et al., 2007), combination of support vector machine and hidden Markov model (Bartlett et al., 2009), syllabification by analogy (SbA) (Adsett and Marchand, 2009), counting the actual syllables to determine the best split of word-medial consonant sequences (Mayer, 2010), segmental conditional random fields (Rogova et al., 2013).

The data-driven approaches give higher accuracy than the rule-based ones for English, a complex syllabic language (Marchand et al., 2007). They also perform better for a language with low syllabic complexity, such as Italian, where the SbA (one of data-driven methods) reaches a word accuracy of 97.70% but the best rule set (SYL-LABE) achieves only 89.77% word accuracy for 44k words (Adsett et al., 2009). Adsett and Marchand proved that data driven approaches generally produce higher accuracy for nine European languages (Adsett and Marchand, 2009), where SbA gives the highest average word accuracy of 96.84% (standard deviation of 2.93) whereas Liang's algorithm produces a mean of 95.67% (standard deviation of 5.70).

Based on the method of classifying languages proposed by Dauer (1983), Indonesian is categorized as a simple syllabic language. It has mostly CV syllables (where C is a single consonant and V is a single vowel) and more open syllables (ended with a vowel) than closed ones (ended with a consonant). A study of 50 k words, collected from the great dictionary of Indonesian language (*Kamus Besar Bahasa Indonesia*, KBBI) created by *Tim Penyusun*

* Corresponding author.
E-mail addresses: suyanto@telkomuniversity.ac.id (S. Suyanto), shartati@ugm.ac.id (S. Hartati), aharjoko@ugm.ac.id (A. Harjoko), Dirk.VanCompernolle@esat.kuleuven.be (D. Van Compernolle).

http://dx.doi.org/10.1016/j.specom.2016.10.009
0167-6393/

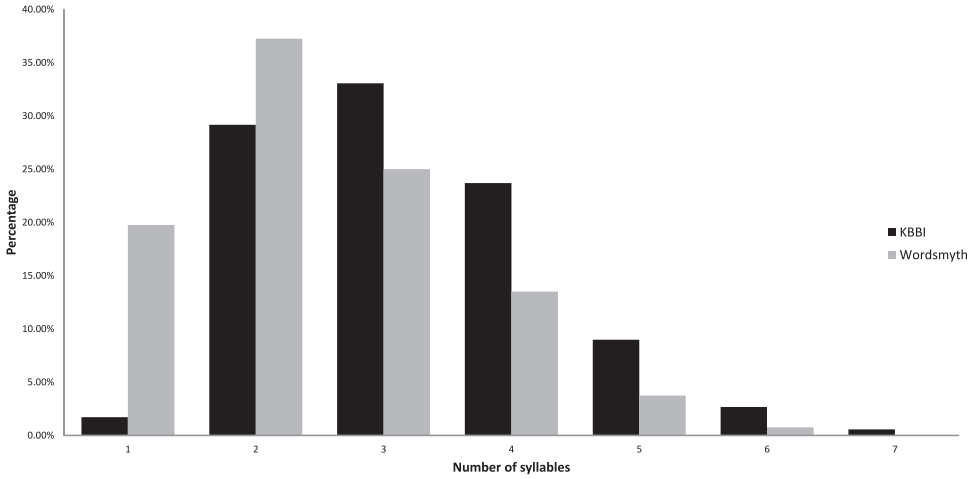

Fig. 1. Distributions of number of syllables per word in the 50 k words from KBBI and Wordsmyth dictionary.

Table 1
Syllable structures in Indonesian language.

Number	Syllable structure	Frequency	Percentage (%)
1	V	6606	4.08
2	CV	82,061	50.63
3	CCV	3056	1.89
4	CCCV	44	0.03
5	VC	6338	3.91
6	CVC	61,826	38.15
7	VCC	116	0.07
8	CVCC	252	0.16
9	CVCCC	6	0.00
10	CCVC	1639	1.01
11	CCVCC	72	0.04
12	CCCVC	56	0.03

Table 2
Examples of the usage of Indonesian affixes.

Root	Affix	Derivative
beli (buy)	*meng-*	*membeli* (buy)
	meng-kan	*membelikan* (buy for)
	per-	*pembeli* (buyer)
	per-an	*pembelian* (purchasing)
	ber-an	*berbelian* (go shopping)
	ter-	*terbeli* (not deliberately bought)
	di-	*dibeli* (deliberately bought)
	di-kan	*dibelikan* (bought by someone)
	-kan	*belikan* (please buy)
	-an	*belian* (purchasing)
henti (stop)	*meng-kan*	*menghentikan* (to stop)
	meng-ber-kan	*memberhentikan* (to stop)
	per-an	*perhentian* (stopping point)
	per-ber-an	*pemberhentian* (stopping point)
	ber-	*berhenti* (to sop)
	ter-	*terhenti* (not deliberately stopped)
	di-kan	*dihentikan* (deliberately stopped)
	-kan	*hentikan* (please stop)

Kamus Pusat Bahasa[1] and syllabified based on the Indonesian syllabification rules developed by Alwi et al. (1998), shows that Indonesian language has 50.63% CV syllables and 56.63% open syllables as listed in Table 1. On the basis of the average syllabic complexity one might conclude that Indonesian is a language with low syllabic complexity. In contrast, English (a complex syllabic language) has 35% of CV syllables and a wider variety of both open and closed syllables (Dauer, 1983). However, Indonesian is a language with mostly long words. Observing the 50 k words of KBBI shows that Indonesian language has 3.20 syllables per word on average (standard deviation 0.41) and 7.54 phonemes per word on average (standard deviation 0.42). It has 98.30% polysyllabic words, much more than monosyllabic words that only account for 1.70%. But, English has 80% polysyllabic words and 20% monosyllabic words based on Wordsmyth dictionary (Marchand et al., 2009). The distributions of the number of syllables per word in 50 k words from KBBI and 50 k words from Wordsmyth dictionary are illustrated by Fig. 1. Meanwhile, the distribution of the number of phonemes per word in the 50 k words from KBBI is shown in Fig. 2.

A data-driven local classifier called PNNR, a variant of *k*-nearest neighbour classification rule (*k*NN), gives a low phoneme error rate (PER) for an Indonesian grapheme-to-phoneme (G2P) conversion (Suyanto and Harjoko, 2014). It is also capable of disambiguating homographs, such as a word '*apel*' where the grapheme ⟨e⟩ could be pronounced as either /ɛ/ like in '*apel pagi*' (morning ceremony) or /ə/ like in '*apel hijau*' (green apple), simply by using a longer graphemic contextual length. In this research, the PNNR is used to develop a new syllabification model for phoneme strings. It receives a phoneme sequence and produces its syllabification points. In this new model, a four-feature phoneme encoding and a phonotactic-based pre-syllabification procedure are proposed. This model will be compared to the Look-Up Procedure (LUP)-based syllabification described by Marchand and colleagues (Marchand et al., 2009), as it may be the closest well established method.

2. Research method

An automatic syllabification is commonly applied to a word (usually called orthographic syllabification) rather than a phoneme sequence (phonemic syllabification). In English, orthographic syllabification is useful to improve the accuracy of G2P. Bartlett et al. proved that the information of orthographic syllabification improves the accuracy of English G2P conversion (Bartlett et al., 2008). However, this fact cannot be generalized to other languages.

Indonesian language has some different characteristics compared to English. It has 29 affixes: 7 prefixes, 4 infixes, and 18 suffixes (Alwi et al., 1998). A prefix or an infix can be used individually or simultaneously with some suffixes to produce derivatives. Two prefixes, with a certain priority order, may be used simultaneously to build a derivative. Therefore, many derivatives can be derived from a root, as in Table 2. These facts make Indonesian language has some ambiguous orthographic syllabifications for

[1] The Lexicographer Team of the Language Center of the Indonesian Ministry of Education and Culture