chiVe 2.0：SudachiとNWJCを用いた 実用的な日本語単語ベクトルの実現に向けて

[1 2] [1] [1] [1]

[1] [3] [3]

[1] [2]

[3]

kawamura_soi@worksap.co.jp

1 はじめに

1 Sudachi Vector chiVe

[*1] chiVe

NWJC 2 Sudachi 3 [*2]

364

Sudachi

Sudachi

3

- : / / / / / / /
- : / / / / /
- : / / /

Sudachi

1

1 Sudachi

	Sudachi		
		N/A	N/A
		N/A	N/A
		/	N/A
		/ / /	/ /

Sudachi

Sudachi A B

C B C

A

NWJC 2 NWJC

1

nwc-toolkit[*3]

NWJC

chiVe NWJC Sudachi A /B /C

1.1 現在のchiVeの問題点

chiVe

4GB

Sudachi

chiVe chiVe 1.0

chiVe chiVe 2.0

[*1] https://github.com/WorksApplications/chiVe

[*2] https://github.com/WorksApplications/Sudachi

[*3] https://code.google.com/archive/p/nwc-toolkit

2 関連研究

2.1 日本語単語ベクトル

2020　1　chiVe 1.0 1

nwjc2vec　4

5　HR

[*4]　hottoSNS-w2v

6　chiVe 1.0　2

2

	MB		
chiVe 1.0	4,171	364	NWJC
nwjc2vec	2,700	155	NWJC
	907	75	
HR	69	17	
hottoSNS-w2v	1,714	206	SNS

2　hottoSNS-w2v

chiVe 1.0　Apache 2.0

2.2 単語ベクトルの圧縮

4　7　8

1)

2)

3)

4)

4

7　8　K

M　K^M

chiVe 1.0

JWSAN-1400 9

M=192/K=64

2GB　chiVe 1.0

3

64

3 chiVe 1.0

	M	K	MB		
chiVe 1.0			4,171MB	54.06	66.53
Shu2018 7	192	64	1,959MB	53.36	66.35
	16	32	212MB	35.36	46.66
	32	16	327MB	32.54	43.64
	64	8	507MB	36.24	48.08

M=16　K=32

98.4%

M=16/K=32　M=32/K=16　M=64/K=8　3

2.3 構成要素からの単語ベクトルの合成

3

chiVe 1.0

Pinter

MIMICK　10

w　v　Bi-LSTM

w　u　u

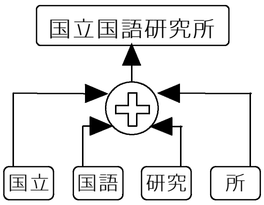
word2vec　fastText　w

MIMICK

11

3 提案手法

Sudachi



1

[*4] https://www.bizreach.co.jp/technology/research/word2vec/

A B C

3.1 短い単位からの長い単位のベクトルを合成

B C A

B C A

B C Sudachi

B C A

B C

A 1 B

C

chiVe 1.0 NWJC

chiVe

2.0

1

4 評価と考察

chiVe 1.0 5

Skip-gram Negative Sampling 12

90

B C A

4.1 B単位・C単位の削減による圧縮の効果

48 A

22.8 B C 3.7 21.6

chiVe 1.0

Sudachi Full

Sudachi Core

BC

7.7% ABC BC 13.9%

4.2 単語類似度・関連度による評価

JWSAN-1400 9

JWSAN-1400 Sudachi

4

4

	52.41	63.69
	50.96	61.67

JWSAN-1400 B C

4.3 文書分類による評価

livedoor [*5]

livedoor 7,367 9

C

10

5

5 livedoor

	0.8478	8.148×10^{-4}
	0.8442	8.649×10^{-4}

1 C

4.4 定性的評価

C

10

6 10

6

1		0.7479		0.8620
2		0.6914		0.8463
3		0.6673		0.8452
4		0.6602		0.7703
5		0.6279		0.7518
6		0.6168		0.7229
7		0.6052		0.6851
8		0.5907		0.6733
9		0.5786		0.6373
10		0.5684		0.6343

A /

2

2

A

[*5] https://www.rondhuit.com/download.html

5 今後の検討

A
B C
A

5.1 よりよい合成方法の検討

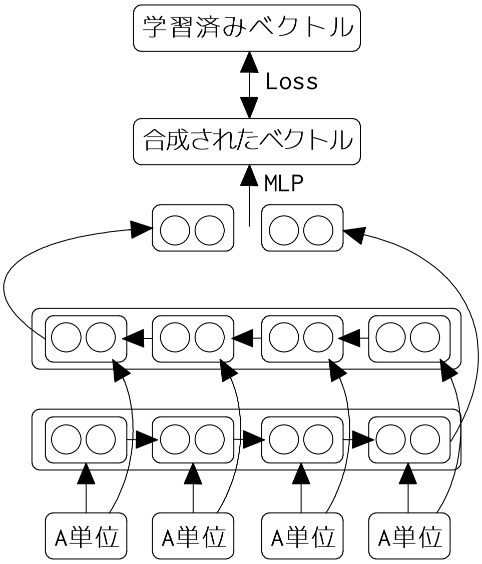
2.3 MIMICK
MIMICK chiVe
2.0 Sudachi 3 A 2
B C A

Sudachi A
B C A

2.2 A

MIMICK
A A

/ / /
A A
A C



2 Bi-LSTM
A

Sudachi C
A

A

5.2 長い単位のベクトルの評価手法の検討

A
B C

6 おわりに

chiVe

Sudachi
A A

ABC

参考文献

1
25 P8-5 2019
2 M. Asahara et al. Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan Alexandria Vol 26 No.1-2 pp.129-148 2014
3 K. Takaoka et al. Sudachi: a Japanese Tokenizer for Business Proceedings of the Eleventh International Conference on Language Resources and Evaluation 2018
4 M. Asahara et al. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus' Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication Vol. 24 No. 2 pp.7-25 2018
5
233 Vol.2017-NL-233 No.17 pp.1-5 2017
6 SNS+Web
2019
7
24 C6-1 2018
8 R.Shu et al. Compressing Word Embeddings via Deep Compositional Code Learning International Conference on Learning Representations 2018
9
24 P10-6 2018.
10 Y. Pinter et al. Mimicking Word Embeddings using Subword RNNs Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing pp.102-112 2017
11
12 T. Mikolov et al. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 2013