

13

Learning 'neat' semantic representations of words

- Neural word embeddings with high usability -

Abstract

A set of word embedding vectors (WEVs) is a “computer-friendly” dictionary that encodes semantic relationships between words into a vector space. It also enables computers to analogically compute semantic relationships among words. Therefore, WEVs have become an important fundamental resource for many natural language processing (NLP) and understanding (NLU) systems, i.e., machine translation, question answering, and communication with robots. We introduce a method for obtaining computer-friendly WEVs. For example, our method can reduce the memory requirement to store WEVs into memory approximately 100 times smaller than those required by conventional methods. As a result, it becomes possible to apply WEVs into applications worked in low-resource devices, such as smartphones. Moreover, our method can also provide high reproducible WEVs whereas WEVs generated by the conventional methods lack such reproducibility.

Word Embedding Vectors

- Representing words by vectors whose dimensions are generally from 50 to 1000
- Important fundamental tool for NLP and NLU
 - NLP: Natural language processing
 - NLU: Natural language understanding

Large text corpus
i.e., news articles
web pages

今日	0.045866	-0.00371	-0.00497	0.070959	-0.08238	0.01114			
明日	-0.3533	-0.03658	0.08013	0.149368	0.297901	-0.09778	-0.14326		
日本	-0.01023	0.075763	0.021459	0.049478	-0.15017	-0.15543	0.222466		
アメリカ	-0.1734	0.003588	-0.09084	0.307067	0.085198	-0.16883	-0.18823		
中国	0.078786	-0.11543	0.02285	-0.06419	0.032716	0.001905	-0.20058		
韓国	-0.02387	-0.05605	-0.20171	-0.15812	-0.14642	-0.04322	-0.36656	-0.31688	0.00000
2016	0.149431	0.079334	0.293562	-0.02928	0.262196	-0.20184	-0.01984	-0.04255	0.111553
円	0.163471	0.032219	-0.03764	-0.00735	0.057619	-0.1467	0.055826	-0.11493	-0.03224
ドル	0.029654	0.105802	-0.32926	-0.10276	-0.27228	-0.11189	0.097063	0.220544	-0.02437
元	-0.16673	0.022848		-0.11159	0.093743	0.062614	-0.00602	0.061345	0.102776
車	0.044333			0.111257	-0.26432	-0.07705	0.277751	-0.10717	-0.2606
									0.201284

Word embedding vectors (WEVs)

Neural networks / Deep learning

Machine translation
Hello! こんにちは

AI
Natural language understanding

Used in many NLP applications

Question answering / Information retrieval
○○ is awesome
weakness: × ×

WEVs generated by conventional methods have some drawbacks when developers apply WEVs to their applications

- Example1: low reproducibility
- Example2: large memory requirement

[high reproducibility] always providing the same computation results under the same condition

[low memory requirement] able to provide WEVs 10 to 100 times smaller than those provides by conventional methods

They encodes some “relationships” between words
⇒ distance and/or angle
enable to compute some “relationships” by vector calculations

Similarity between words
⇒ distance

i.e., words in similar meaning will be located closely in the vectors space

Certain relationship
⇒ angle

i.e., singular tense and the plural tense (bleu)

References

- [1] J. Suzuki, M. Nagata, “Right-truncatable neural word embeddings,” in *Proc. of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1145-1151, 2016.
- [2] J. Suzuki, M. Nagata, “Learning compact neural word embeddings by parameter space sharing,” in *Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2046-2052, 2016.

Contact

Jun Suzuki Linguistic Intelligence Research Group, Innovative Communication Laboratory
Email : suzuki.jun(at)lab.ntt.co.jp