Modified Multi-Sense Skip-Gram using weighted context and X-means

가중된 문맥과 X-means방법을 이용한 수정된 다중 의미 스킵 그램

Hyunwoo Jeong, Eun Ryung Lee

Department Statistics, Sungkyunkwan University

May, 2021

1 Introduction

2 Proposed Methodology

3 Simulation

Word Sense Disambiguation

Word Sense Disambiguation, WSD

"There is a nuclear **plant** near the forest."
"All **plants** need light and water"

1. In NLP, the main method is to assign only one vector per word, assuming a single meaning (Skip-gram).
2. It is difficult to understand the meaning of two sentences properly if you judge multi-sense words in one meaning.
3. This problems can be solved by a Multi-Sense Skip-Gram (MSSG) method that assigns vectors to each meaning of a Multi-sense word.

Multi-Sense Skip-Gram(MSSG) (Neelakantan, 2014) (1)

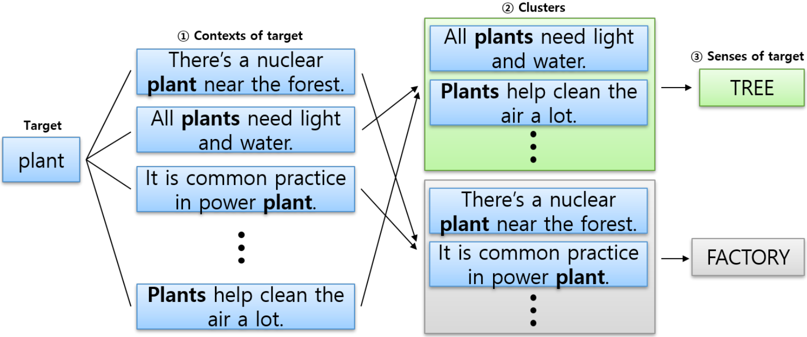

Figure 1: MSSG model's idea

Multi-Sense Skip-Gram(MSSG) (Neelakantan, 2014) (2)

1. Generating context vector
   - $v_{context}(C_t) = 1/(2R_m) \times \sum_{c \in C_t} v_g(c)$
   - Using embedding vector of Skip-Gram as global vector
2. Determine sense of w_t: By conducting K-means clustering
3. Update Senses vector
   - $\mu(w, k) = v_s(w, k)$
   - Update sense vector using Skip-Gram's negative sampling method

Multi-Sense Skip-Gram(MSSG) (Neelakantan, 2014) (3)

Definition

- T: number of context target word w appears
- w_t: target word of t-th context
- C_t: t-th context words set for w
- R_m: window size
- $v_g(c) \in R^d$: global embedding vector for c
- $\mu(w, k)$: k-th cluster center for word w
- s_t: sense for w_t
- $v_s(w, k)$: k-th sense vector for word w

Problems

Problems

1. Biased meaning of sense vectors
   - "Nuclear power **plants** can be harmful to trees."
   - "Plant habitats are large near power facilities."
2. Fixed number of senses for target word
   - By using K-means clustering

Solution

1. We propose a weighted context vector method with different importance to the location of the context words.
2. We use X-means clustering to divided contexts by meaning and estimate the number of meanings.

1. Introduction
2. Proposed Methodology
3. Simulation

Weighted context vector

Weighted context vector

$$v_{\mathrm{w,cont}}(C_t) = \sum_{|k|=1,2,\ldots,R} \mathrm{d}_k \times v_g(c_{t+k}) \qquad (1)$$

- $C_t = \{c_{t-R_m}, \ldots, c_{t-1}, c_{t+1}, \ldots, c_{t+R_m}\}$
- $d_k > 0$, $\sum_{|k|=1,2,\ldots,R} \mathrm{d}_k = 1$
- d_k: monotone decreasing with $|k|$ values $k = \pm 1, 2, \ldots, R$

In this study, we consider weight d_k that linearly decrease with $|k|$ values.

Weighted context vector for 2 sentences

1. "Nuclear power **plants** can be harmful to trees."

$$\Rightarrow v_{w,cont(C_{\mathbf{1}})} = \frac{4 \times v_g(nuclear)}{24} + \frac{5 \times v_g(power)}{24} + \frac{5 \times v_g(can)}{24} + \cdots + \frac{1 \times v_g(trees)}{24}$$

2. "Plant habitats are large near power facilities."

$$\Rightarrow v_{w,cont(C_{\mathbf{1}})} = \frac{6 \times v_g(habitat)}{24} + \frac{5 \times v_g(are)}{24} + \cdots + \frac{2 \times v_g(power)}{24} + \frac{1 \times v_g(facilities)}{24}$$

X-means clustering (Dan, Moore 2000)

X-means clustering

- The X-means clustering method performs clustering based on the K-means clustering.
- Also, optimizes the number of clusters based on the BIC measure assuming the distribution of the data as normal

X-means clustering algorithm

1. For a given initial number of cluster K_0(mainly $K_0 = 2$), perform K-means clustering.
2. As a result of step1, for K_0 clusters $S_1, \ldots, S_{K_0}$, the following processes are carried out.
   - A For cluster S_k, perform 2-means clustering.
   - B. For $S_k^{(1)}, S_k^{(2)}$ the result of 2-1, assume a normal dist. and compute BIC
     - a. If $\mathrm{BIC}(S_k) > \mathrm{BIC}(S_k^{(1)}, S_k^{(2)})$, determine $(S_k^{(1)}, S_k^{(2)})$ and perform A $\sim$ C
     - b. If $\mathrm{BIC}(S_k) < \mathrm{BIC}(S_k^{(1)}, S_k^{(2)})$, determine S_k
   - C. Complete split for cluster S_k and re-number the final clusters of S_k

1. Introduction
2. Proposed Methodology
3. Simulation

Data description

Data description

- Corpus: 584 abstracts in Journal of Statistical Software(JSS)
- 5,353 vocabulary and 45,106 word tokens
- target word ‘plant’ occures 72 times / 32 ‘tree’ and 40 ‘facility’

Preprocessing

- Lemmatization, stopword processing by spacy library
- Exclude words with length 1 (’a’, ’R’, etc.)

Global Vector Extraction

- By using Skip-Gram method
- embedding vector size: 300 $\sim$ 500 by 100
- Window: 1 $\sim$ 9 by 2

Simulation Result (1)

Analysis for context vector methods

- By performing 2-means clustering, we express the 10 words closest(similar) and similarity value.

Method	Sense	5 Nearest Neighbor Words
$v_g(plant)$	-	gas(0.50) nuclear(0.45) grf(0.44) operational(0.43)
Unweighted	tree	seed(0.74) trigger(**0.72**) growth(0.68) fertilization(0.66)
	facility	power(0.70) nuclear(0.60) gas(**0.57**) energy(**0.53**)
Weighted	tree	seed(**0.76**) trigger(**0.72**) growth(**0.72**) fertilization(**0.67**)
	facility	power(**0.79**) nuclear(**0.61**) gas(**0.57**) energy(**0.53**)

Table 1: 4 Nearest neighbor words for each context vector

Simulation Result (2)

Analysis cluster number estimation and classification accuracy

- (Left) cluster number estimation for X-means clustering
- (Right) classification accuracy for context vector method
- embedding size: 300, window: 5, iteration: 1000

	Cluster number					
R_m	2	3	4	5	6	7
1	665	249	62	19	4	1
3	645	260	71	20	3	1
5	999	1	0	0	0	0
7	1000	0	0	0	0	0
9	1000	0	0	0	0	0

	Proposed		Ordinary	
clustering	mean	median	mean	median
X-means	**87.7**	**88.9**	85.8	87.5
2-means	**88.2**	**88.9**	88.0	**88.9**
3-means	**74.3**	73.6	73.6	**75.0**
4-means	63.4	**63.9**	**64.1**	62.5
5-means	55.6	55.6	**55.8**	**56.9**

Simulation result (3)

- Compare classification accuracy for MSSG and Modified MSSG
- iteration: 100

Method	d	$R_m = 1$	3	5	7	9
MSSG	300	79.8	87.8	**90.1**	88.1	89.5
Modified MSSG	300	78.3	79.3	**80.8**	74.3	59.8
MSSG	400	81.3	85.8	**88.2**	87.1	88.0
Modified MSSG	400	78.0	75.5	**89.1**	89.1	89.0
MSSG	500	79.8	88.6	87.1	87.5	**89.4**
Modified MSSG	500	74.0	84.7	88.4	**89.2**	88.7

Conclusion

- At well-specified parameters (R_m, d), Modified MSSG accurately estimate the number of senses of the target word.
- Also, Modified MSSG perform similarly or rather well as the embedding performance of MSSG using the true number of clusters.
- Therefore, we demonstrate significant improvements by providing improved semantic accuracy of embiedding vectors and efficiency for estimating sense numbers.

Appendix

Skip-Gram (Mikolov, 2013)

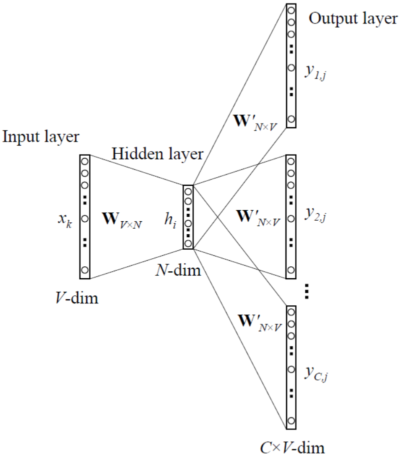

Figure 2: Structure of Skip-gram

Appendix

- given a pair of words (w_t, c), the probability that word c is observed in the context of target word w_t is

$$P(D = 1|v(w_t), v(c)) = \frac{e^{v(c)^T v(w_t)}}{\sum e^{v(w_t)^T v(c)}}$$

- the probability of not observing word c in the context of target word w_t is

$$P(D = 0|v(w_t), v(c)) = 1 - P(D = 1|v(w_t), v(c))$$

Appendix

Negative sampling

- word embeddings are learned by maximizing the objective function:

$$J(\theta) = \sum_{(w_t, c_t) \in D^+} \sum_{c \in C_t} logP(D = 1 | v(w_t), v(c)) + \sum_{(w_t, c_t') \in D^-} \sum_{c' \in C_t} logP(D = 0 | v(w_t), v(c'))$$

- where c_t' t is randomly sampled noisy context words for word w_t.

The End