CS 6501 Natural Language Processing

Word Embeddings

Yangfeng Ji

October 24, 2018

Department of Computer Science
University of Virginia

UNIVERSITY of VIRGINIA | ENGINEERING

Overview

1. Theoretical Framework
2. Evaluation Methods
3. Problems

Theoretical Framework

Last Lecture

$$\langle \boldsymbol{u}_{w_j}, \boldsymbol{v}_{w_i} \rangle \approx \log(X_{ij}) + g(\mathbf{X}) \tag{1}$$

A Generative Model with Random Walk

$$P(w_t \mid c_t) \propto \exp(\langle \boldsymbol{c}_t, \boldsymbol{v}_{w_t} \rangle), \tag{2}$$

where $\boldsymbol{c}_t, \boldsymbol{v}_{w_t}$ are the numeric representations of discourse c_t and word w_t.

- $\boldsymbol{c}_t$ does a slow random walk on the unit sphere
- $\boldsymbol{c}_{t+1}$ is obtained from $\boldsymbol{c}_t$ by adding a small random displacement vector
- Intuition: the topic of a coherent text should be continuous mostly

[Arora et al., 2016]

Partition Function Z

Aka, normalization term Z

$$P(w_t \mid c_t) = \frac{1}{Z_t} \exp(\langle \boldsymbol{c}_t, \boldsymbol{v}_{w_t} \rangle), \tag{3}$$

where

$$Z_t = \sum_{w'} \exp(\langle \boldsymbol{c}_t, \boldsymbol{v}_{w'} \rangle) \tag{4}$$

Partition Function Z

Aka, normalization term Z

$$P(w_t \mid c_t) = \frac{1}{Z_t} \exp(\langle \boldsymbol{c}_t, \boldsymbol{v}_{w_t} \rangle), \tag{3}$$

where

$$Z_t = \sum_{w'} \exp(\langle \boldsymbol{c}_t, \boldsymbol{v}_{w'} \rangle) \tag{4}$$

Usually,

$$Z_t \neq Z_{t'} \tag{5}$$

where $t \neq t'$

Partition Function Z (II)

One additional assumptions

- $v = s \cdot \hat{v}$, $\hat{v}$ is from the spherical Gaussian distribution and s is a scalar random variable
- (Previous assumption) $\boldsymbol{c}_t$ does a slow random walk on the unit sphere

Partition Function Z (II)

One additional assumptions

- $\boldsymbol{v} = s \cdot \hat{\boldsymbol{v}}$, $\hat{\boldsymbol{v}}$ is from the spherical Gaussian distribution and s is a scalar random variable
- (Previous assumption) $\boldsymbol{c}_t$ does a slow random walk on the unit sphere

Self-normalization effect [Andreas and Klein, 2015]

$$Z_t \approx Z_{t'} \tag{6}$$

$P(w, w')$

Consider two adjacent words w, w' in text, we have

$$\begin{aligned} P(w, w') &= E_{c,c'}[P(w, w' \mid c, c')] && (7) \\ &= E_{c,c'}[P(w \mid c)P(w' \mid c')] && (8) \\ &= E_{c,c'}\left[\frac{\exp(\langle \boldsymbol{c}, \boldsymbol{v}_w \rangle)}{Z} \cdot \frac{\exp(\langle \boldsymbol{c}', \boldsymbol{v}_{w'} \rangle)}{Z'}\right] && (9) \end{aligned}$$

$P(w, w')$

To continue the derivation, we need to know

1. $Z \approx Z'$
2. $P(c' \mid c)$ characterizes the transition from $\boldsymbol{c}$ to $\boldsymbol{c}'$ in random walk
3. central limit theorem

Conclusion'

$$\log P(w, w') = \frac{\|\boldsymbol{v}_w + \boldsymbol{v}_{w'}\|_2^2}{2d} - 2\log Z \pm \epsilon \tag{10}$$

$P(w, w')$ (II)

$$\log P(w, w') = \frac{\|\boldsymbol{v}_w + \boldsymbol{v}_{w'}\|_2^2}{2d} - 2\log Z \pm \epsilon \tag{11}$$

$$\log P(w) = \frac{\|\boldsymbol{v}_w\|_2^2}{2d} - \log Z \pm \epsilon \tag{12}$$

$$\mathrm{PMI}(w, w') = \frac{\langle \boldsymbol{v}_w, \boldsymbol{v}_{w'} \rangle}{d} \pm \mathcal{O}(\epsilon) \tag{13}$$

Evaluation Methods

Overview

- Intrinsic Evaluation
  - Word similarity
  - Word analogy
  - Word intrusion
- Extrinsic Evaluation

Word Similarity

Let w_i and w_j be two words, and $\boldsymbol{v}_{w_i}$ and $\boldsymbol{v}_{w_j}$ be the corresponding word embeddings, word similarity can be obtained by computing their cosine similarity between $\boldsymbol{v}_{w_i}$ and $\boldsymbol{v}_{w_j}$ as

$$\cos(\boldsymbol{v}_{w_i}, \boldsymbol{v}_{w_j}) = \frac{\langle \boldsymbol{v}_{w_i}, \boldsymbol{v}_{w_j} \rangle}{\|\boldsymbol{v}_{w_i}\| \cdot \|\boldsymbol{v}_{w_j}\|} \tag{14}$$

where $\langle \cdot, \cdot \rangle$ is the inner product of two vectors, $\| \cdot \|$ is the ℓ_2 norm of a vector.

Examples

Word_1	Word_2	Similarity score [0,10]
love	sex	6.77
stock	jaguar	0.92
money	cash	9.15
development	issue	3.97
lad	brother	4.46

Figure: Sample word pairs along with their human similarity judgment from WS-353 [Faruqui et al., 2016].

Datasets

Available word similarity datasets

Dataset	Word pairs	Reference
RG	65	Rubenstein and Goodenough (1965)
MC	30	Miller and Charles (1991)
WS-353	353	Finkelstein et al. (2002)
YP-130	130	Yang and Powers (2006)
MTurk-287	287	Radinsky et al. (2011)
MTurk-771	771	Halawi et al. (2012)
MEN	3000	Bruni et al. (2012)
RW	2034	Luong et al. (2013)
Verb	144	Baker et al. (2014)
SimLex	999	Hill et al. (2014)

Figure: Word similarity datasets [Faruqui et al., 2016].

Word Similarity

the basis for other intrinsic evaluation

Word Similarity: Problems (I)

Similarity and relatedness: which pair is closer?

- `train, car`
- `coffee, cup`

Word Similarity: Problems (I)

Similarity and relatedness: which pair is closer?

- `train, car`
- `coffee, cup`

In WS-353, the similarity between `coffee` and `cup` is higher than `train` and `car`.

Word Similarity: Problems (II)

Frequency effects of cosine similarity

- prevents the bias introduced by the norm of a vector 3
- pairs of words that have similar frequency will be closer in the embedding space
  - higher similarity of two words can be given by cosine similarity then they should be based on their word meaning

Word Similarity: Problems (III)

Inability to account for polysemy (one word has multiple meanings)

$$\cos(\boldsymbol{v}_{w_i}, \boldsymbol{v}_{w_j}) = \frac{\langle \boldsymbol{v}_{w_i}, \boldsymbol{v}_{w_j} \rangle}{\|\boldsymbol{v}_{w_i}\| \cdot \|\boldsymbol{v}_{w_j}\|} \tag{15}$$

- Encode them into different dimensions?

Word Analogy

- It is sometimes referred as *linguistic regularity* [Mikolov et al., 2013]
- The basic setup

$$w_a : w_b = w_c :?$$

  where $w_{a,b,c}$ are words and w_a, w_b are related under a certain linguistic relation
- Calculation: $\langle \boldsymbol{v}_{w_a} - \boldsymbol{v}_{w_b}, \boldsymbol{v}_{w_c} - \boldsymbol{v}_{w_d} \rangle$
- Example
  - Semantic `love` : `like`
  - Syntactic `quick` : `quickly`
  - Gender `king` : `man`
  - Others `Beijing` : `China`

Word Analogy: Examples

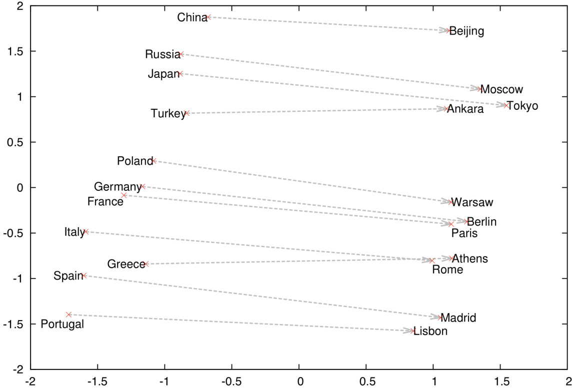


Figure: Word analogy examples.

Word intrusion

From [Faruqui et al., 2014]

`naval, industrial, technological, marine, identity`

- constructed from word embeddings
- evaluated by human annoators

Extrinsic Evaluation

- Implicit assumption: there is a consistant, global ranking of word embedding quality, and that higher quality embeddings will necessarily improve results on *any* downstream task.
- Unfortunately, this assumption does not hold in general [Schnabel et al., 2015].
- Examples
  - empirical results show that it may not be able give much help to syntactic parsing [Andreas and Klein, 2014]
  - adding surface-form features always help ([Ji and Eisenstein, 2014a] and many other works)

Problems

Gender Bias

$$v_{\texttt{man}} - v_{\texttt{woman}} \approx v_{\texttt{computer programmer}} - v_{\texttt{homemaker}} \quad (16)$$

$$v_{\texttt{father}} - v_{\texttt{mother}} \approx v_{\texttt{doctor}} - v_{\texttt{nurse}} \quad (17)$$

[Bolukbasi et al., 2016]

Problem

Word embeddings like this not only reflect such stereotypes but also amplify them

A Solution

Three steps [Bolukbasi et al., 2016]

1. find gender neutral words with biases in the original embeddings;
2. identify the gender-specific space V and its orthogonal complement $V^{\perp}$
3. project embeddings of the gender neutral words to the subspace $V^{\perp}$

Example

tote reading records clip commit game
browsing sites seconds slow arrival tactical
crafts credits
trimester tanning user parts drop reel firepower
ultrasound busy hoped command
housing caused ill scrimmage
modeling beautiful oils self gel looks zeal builder drafted
sewing dress dance steals effect trips brilliant
flirt nuclear yard genius
pageant earrings divorce firms cocky journeyman
nurses seeking ties guru
salon tearful cold buddy
cow voters youth
sassy breasts pearls vases iv regional rule
firmly buddies burly
homemaker dancer friend
babe lamb folks priest mate beard
mommy witch witches dads boys cousin chap boyhood
actresses gals wives lad
fiance sons son
queen girlfriends girlfriend brothers
sisters grandmother wife daddy nephew
ladies fiancee
daughters
biased
he
okay

Question

Can we have an interpretability of each dimension?

- like what we have from topic models (e.g., Latent Dirichlet Allocation)?

Solution

Post-processing on word embeddings

- restructing with sparsity constraint [Faruqui et al., 2015]
- rotating word embedding space using factor analysis [Park et al., 2017]

Reconstruction with Sparsity

Interpretability is *derived* from the sparsity constraint as

$$\arg\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{V} \|\boldsymbol{x}_i - \mathbf{D}\boldsymbol{a}_i\|_2^2 + \lambda\|\boldsymbol{a}_i\|_1 + \tau\|\mathbf{D}\|_2^2 \tag{18}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$ are the original and sparse embeddings of word i, $\mathbf{D}$ is the transformation matrix.

Example

X	combat, guard, honor, bow, trim, naval 'll, could, faced, lacking, seriously, scored see, n't, recommended, depending, part due, positive, equal, focus, respect, better sergeant, comments, critics, she, videos
A	fracture, breathing, wound, tissue, relief relationships, connections, identity, relations files, bills, titles, collections, poems, songs naval, industrial, technological, marine stadium, belt, championship, toll, ride, coach

Figure: Top-ranked words per-dimension before and after reconstruction. Each line shows words from a different dimension.

Problem

- Word embeddings from either Word2vec or GloVe encode not just semantic information

Problem

- Word embeddings from either Word2vec or GloVe encode not just semantic information
- In some applications, we want to emphasize one particular aspect of linguistic information
  - Semantic information [Faruqui et al., 2014]
  - Discourse information [Ji and Eisenstein, 2014b]

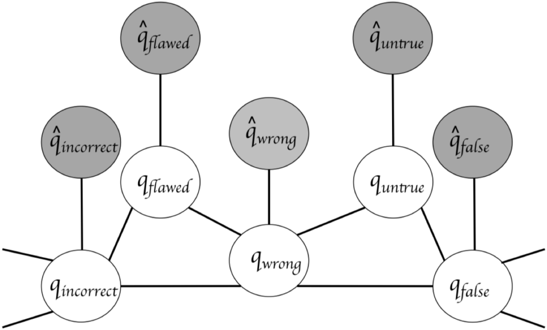
Problem

- Word embeddings from either Word2vec or GloVe encode not just semantic information
- In some applications, we want to emphasize one particular aspect of linguistic information
  - Semantic information [Faruqui et al., 2014]
  - Discourse information [Ji and Eisenstein, 2014b]
- Solutions
  - retrofitting word embeddings [Faruqui et al., 2014]
  - learning from supervision information [Ji and Eisenstein, 2014b]

Retrofitting

Retrofitting with WordNet [Miller, 1995]

- $\Omega = (V, E)$ be a semantic graph over words, where V is the node set with each element as a word, and E is the edge set with each edge representing a semantic relation between two words.



Retrofitting (II)

- The goal is to learn word embeddings $\{\tilde{v}\}$ such that $\tilde{\boldsymbol{v}}_i$ and $\tilde{\boldsymbol{v}}_j$ are close enough if $(i, j) \in E$.
- In addition, $\{\tilde{v}\}$ should also statisfy the constraint from original word embeddings, such that $\tilde{v}_i$ and $\tilde{\boldsymbol{v}}_i$ are close enough for every word in $\mathcal{V}$.

$$\Psi(\tilde{\mathbf{V}}) = \sum_{i=1}^{|\mathcal{V}|} \left[\alpha_i \|\boldsymbol{v}_i - \tilde{\boldsymbol{v}}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|\tilde{\boldsymbol{v}}_i - \tilde{\boldsymbol{v}}_j\|^2 \right] \quad (19)$$

Learning from Supervision Signal

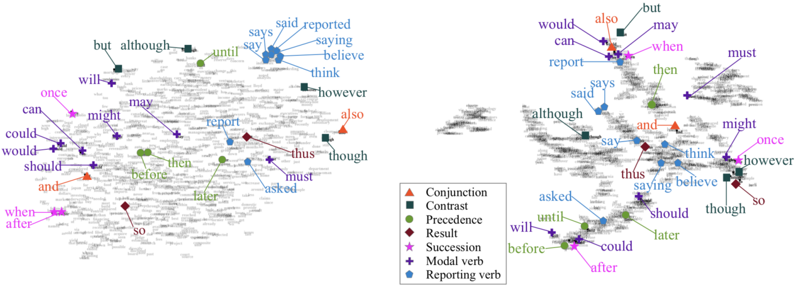


Figure: (Left) Word embeddings learned with supervision signal; (Right) Unsupervised word embeddings.

Summary

Reference

Andreas, J. and Klein, D. (2014).
How much do word embeddings encode about syntax?
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 822–827.

Andreas, J. and Klein, D. (2015).
When and why are log-linear models self-normalizing?
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–249.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016).
A latent variable model approach to pmi-based word embeddings.
Transactions of the Association for Computational Linguistics, 4:385–399.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016).
Man is to computer programmer as woman is to homemaker? debiasing word embeddings.
In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2014).
Retrofitting word vectors to semantic lexicons.
arXiv preprint arXiv:1411.4166.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016).
Problems with evaluation of word embeddings using word similarity tasks.
arXiv preprint arXiv:1605.02276.

Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., and Smith, N. (2015).
Sparse overcomplete word vector representations.
arXiv preprint arXiv:1506.02004.

Ji, Y. and Eisenstein, J. (2014a).