Investigating Word-Class Distributions in Word Vector Spaces

Ryohei Sasano[†] Anna Korhonen[‡]

† Graduate School of Informatics, Nagoya University, Japan

‡ Language Technology Lab, University of Cambridge, UK

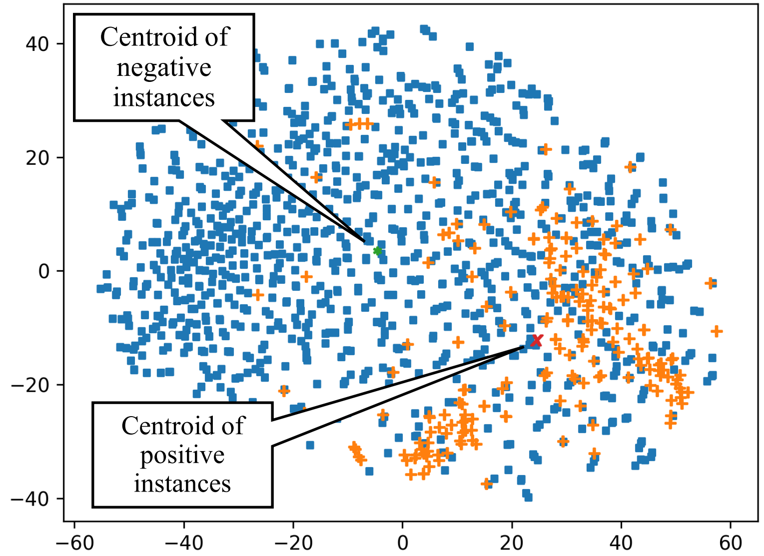
Word-Class in a Word Vector Space

- Many successes in representing word meanings with a vector (e.g.,. CBOW, skip-gram, GloVe)
- Their interpretation and geometry have also attracted attention [Kim+'13, Mimno+'17]

- Little attention has been paid to the distribution of words belonging to a certain word class
  e.g., Semantic class of direct objects of verb *play*
  - +: words that can be a direct object (positive instances)
  - ■: the other words (negative instances)
  - Positive instances tend to be densely distributed around their centroid
  - but not evenly distributed near the centroid

⇩

- Investigate word-class distributions in word vector spaces

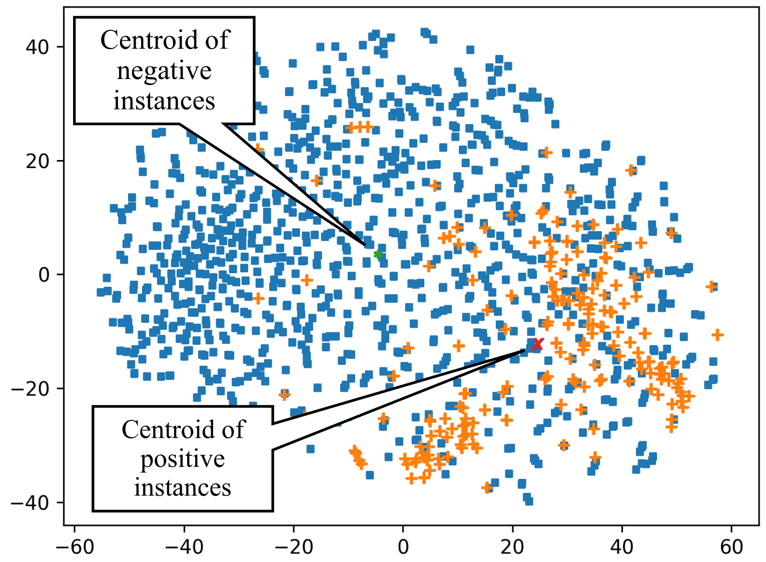


2D t-SNE projection of GloVe vectors

Word-Class in a Word Vector Space

- Many successes in representing word meanings with a vector (e.g.,. CBOW, skip-gram, GloVe)
- Their interpretation and geometry have also attracted attention [Kim+'13, Mimno+'17]

- Little attention has been paid to the distribution of words belonging to a certain word class
  e.g., Semantic class of direct objects of verb *play*
    - +: words that can be a direct object (positive instances)
    - ■: the other words (negative instances)
  - Positive instances tend to be densely distributed around their centroid
  - but not evenly distributed near the centroid



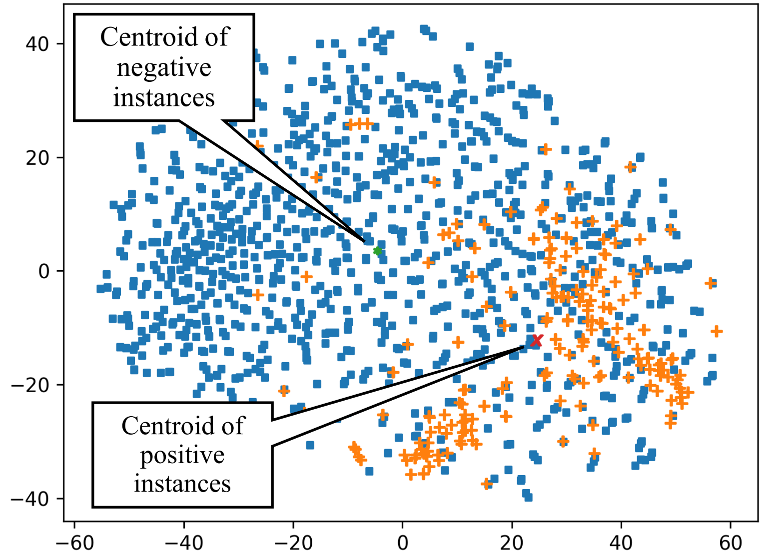
2D t-SNE projection of GloVe vectors

- Investigate word-class distributions in word vector spaces

Word-Class in a Word Vector Space

- Many successes in representing word meanings with a vector (e.g.,. CBOW, skip-gram, GloVe)
- Their interpretation and geometry have also attracted attention [Kim+'13, Mimno+'17]

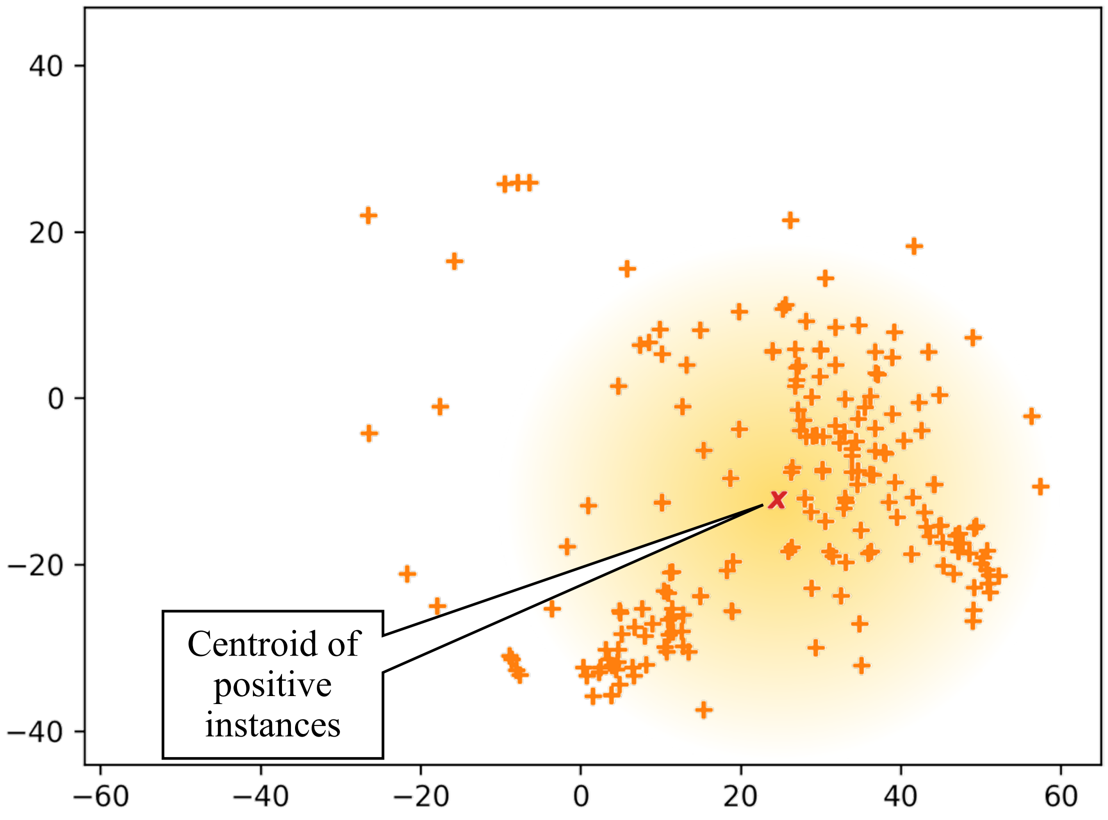
- Little attention has been paid to the distribution of words belonging to a certain word class
  e.g., Semantic class of direct objects of verb *play*
  - +: words that can be a direct object (positive instances)
  - ■: the other words (negative instances)
  - Positive instances tend to be densely distributed around their centroid
  - but not evenly distributed near the centroid



2D t-SNE projection of GloVe vectors

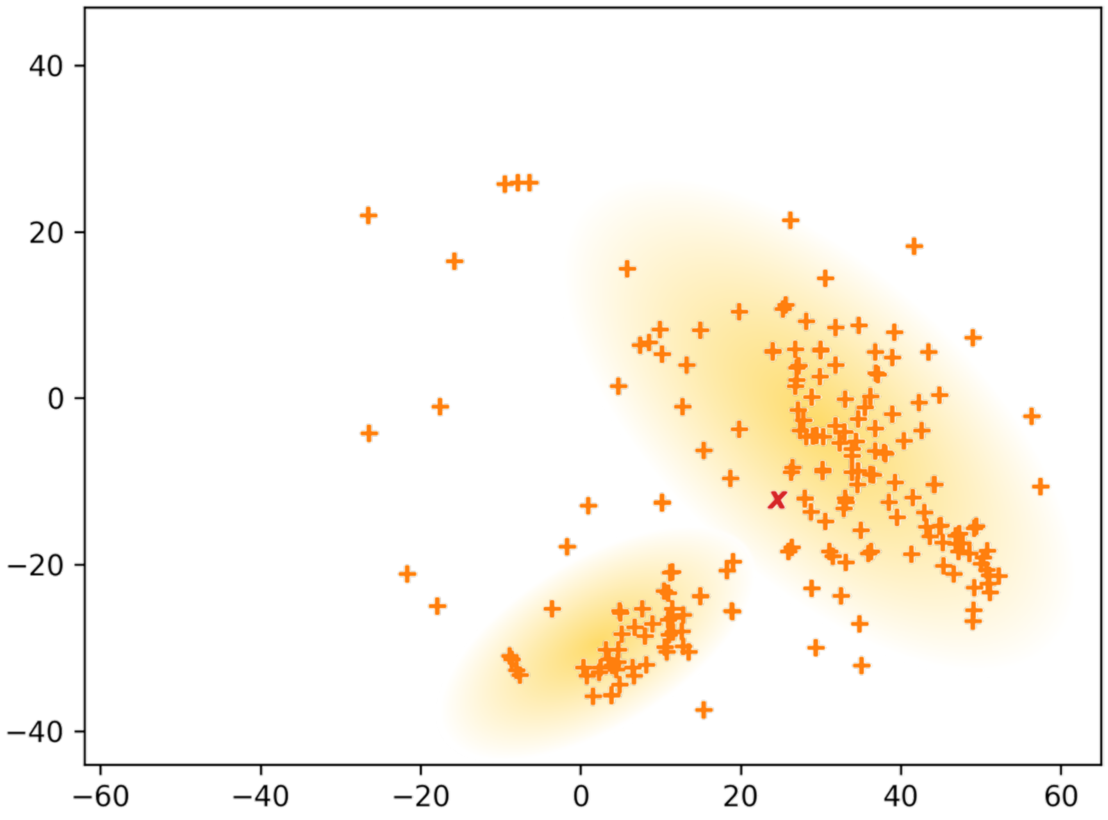
- Investigate word-class distributions in word vector spaces

How are words belonging to a word class distributed in the word vector spaces?



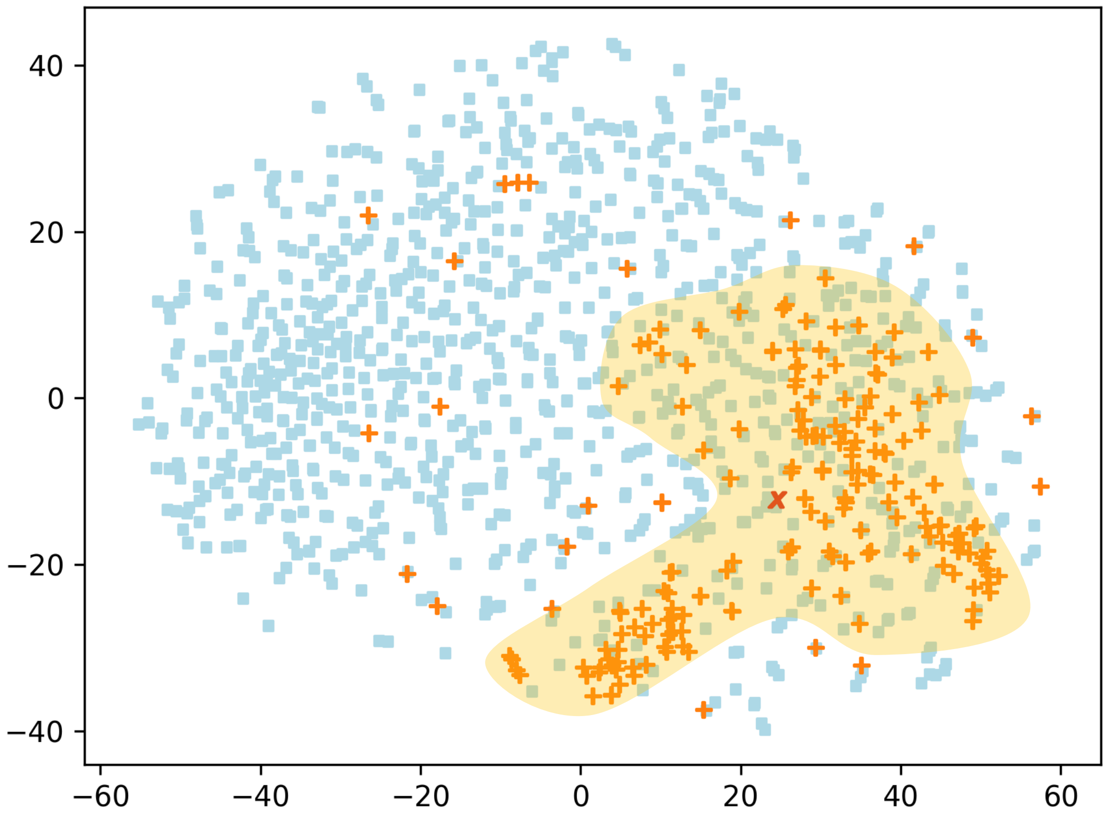
1. Can a simple centroid-based approach provide a reasonably good model?
2. Is it useful to consider the geometry of the distribution and the existence of subgroups for modeling the distribution
3. Is it essential to consider the negative instances to achieve adequate modeling?

How are words belonging to a word class distributed in the word vector spaces?



1. Can a simple centroid-based approach provide a reasonably good model?
2. Is it useful to consider the geometry of the distribution and the existence of subgroups for modeling the distribution
3. Is it essential to consider the negative instances to achieve adequate modeling?

How are words belonging to a word class distributed in the word vector spaces?



1. Can a simple centroid-based approach provide a reasonably good model?
2. Is it useful to consider the geometry of the distribution and the existence of subgroups for modeling the distribution
3. Is it essential to consider the negative instances to achieve adequate modeling?

Our Approach

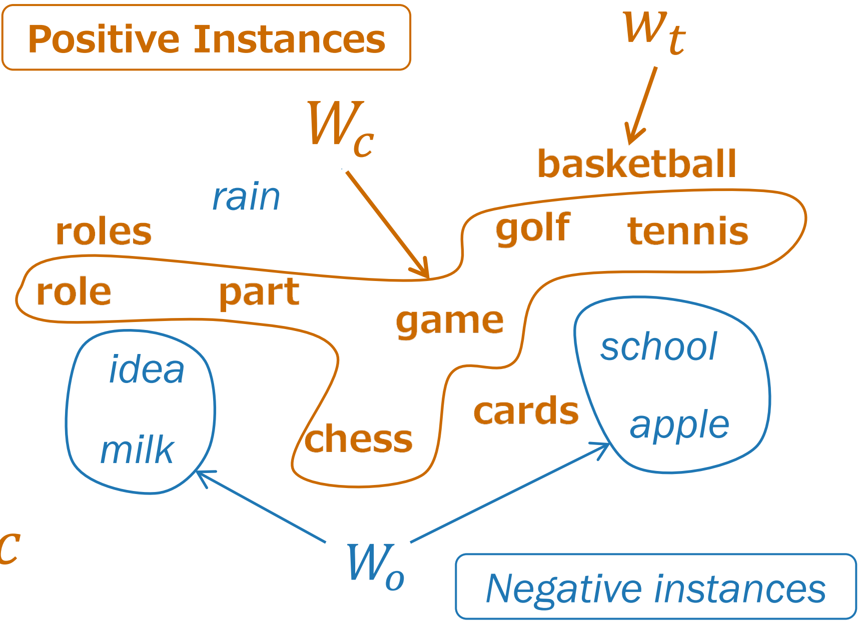
1. Make several assumptions about the distribution

2. Model the distribution accordingly

3. Validate each assumption by comparing the goodness of each model

Problem formulation

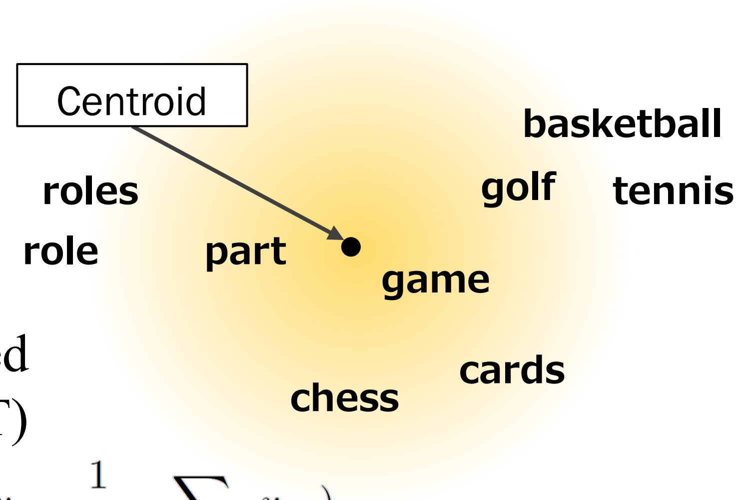
- Notation
  - c: word class (e.g., direct objects of verb *play*)
  - W_c: subset of words that belong to c
  - w_t: target word that can be a member of c but is not included in W_c
  - W_o: subset of words that *do not* belong to c



- Objective
  - Find a scoring function $\mathrm{f}(w, W_c)$ that assigns a higher score to w_t and lower scores to $w_o \in W_o$ (e.g., higher score to *basketball* than to *idea*, *milk*, *school*, *apple*)

Models

5 Models without negative instances



(1) Centroid-based model (CENT)

$$f_{\mathrm{CENT}}(w, W_c) = \cos(v_w, \frac{1}{|W_c|} \sum_{w_c \in W_c} v_{w_c})$$

(2) Gaussian model (GM)

$$f_{\mathrm{GM}}(w, W_c) = \mathcal{N}(v_w | \mu, \Sigma)$$

(3) Gaussian mixture model (GMM)

$$f_{\mathrm{GMM}}(w, W_c) = \sum_{k=1}^{K} \pi_k \mathcal{N}(v_w | \mu_k, \Sigma_k)$$

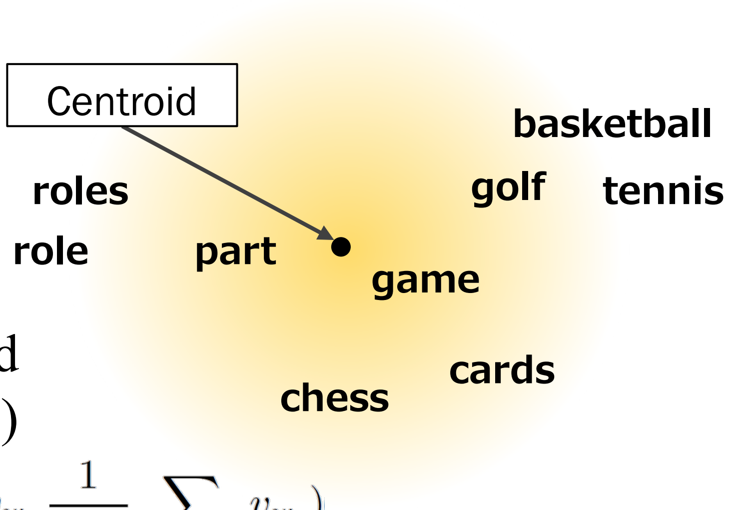
(4) k-nearest neighbor model (GM)

$$f_{k\mathrm{NN}}(w, W_c) = \frac{1}{k} \sum_{w_c \in k\mathrm{NN}_w(W_c)} \cos(v_w, v_{w_c})$$

(5) One-class SVM (1-SVM)

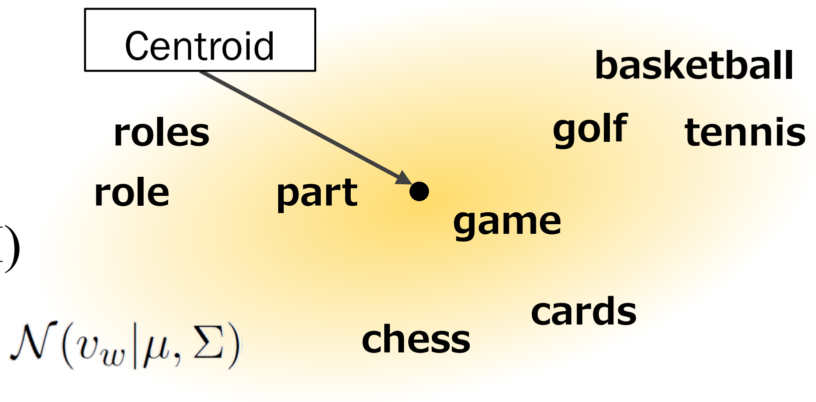
5 Models without negative instances

(1) Centroid-based model (CENT)



$$\mathrm{f}_{\mathrm{CENT}}(w, W_c) = \cos(v_w, \frac{1}{|W_c|}\sum_{w_c \in W_c} v_{w_c})$$

(2) Gaussian model (GM)



$$\mathrm{f}_{\mathrm{GM}}(w, W_c) = \mathcal{N}(v_w|\mu, \Sigma)$$

(3) Gaussian mixture model (GMM)

$$\mathrm{f}_{\mathrm{GMM}}(w, W_c) = \sum_{k=1}^{K} \pi_k \mathcal{N}(v_w|\mu_k, \Sigma_k)$$

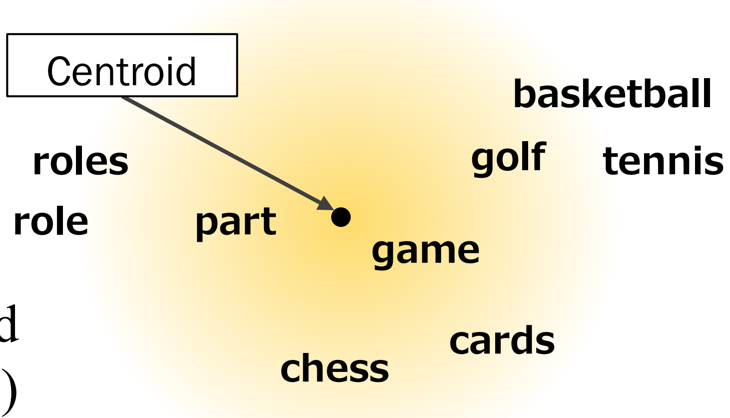
(4) k-nearest neighbor model (GM)

$$\mathrm{f}_{k\mathrm{NN}}(w, W_c) = \frac{1}{k}\sum_{w_c \in k\mathrm{NN}_w(W_c)} \cos(v_w, v_{w_c})$$

(5) One-class SVM (1-SVM)

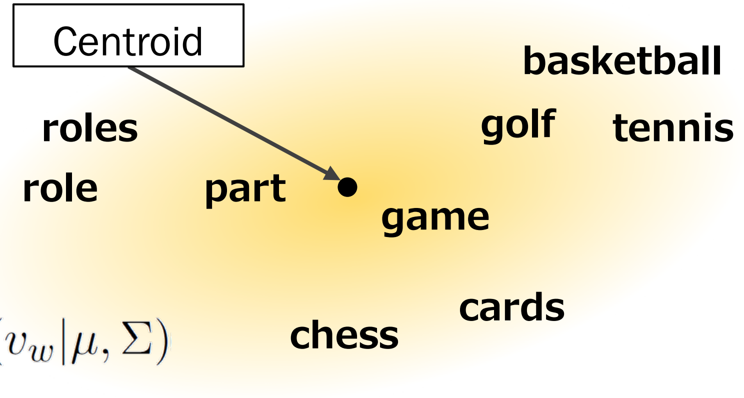
5 Models without negative instances

(1) Centroid-based model (CENT)



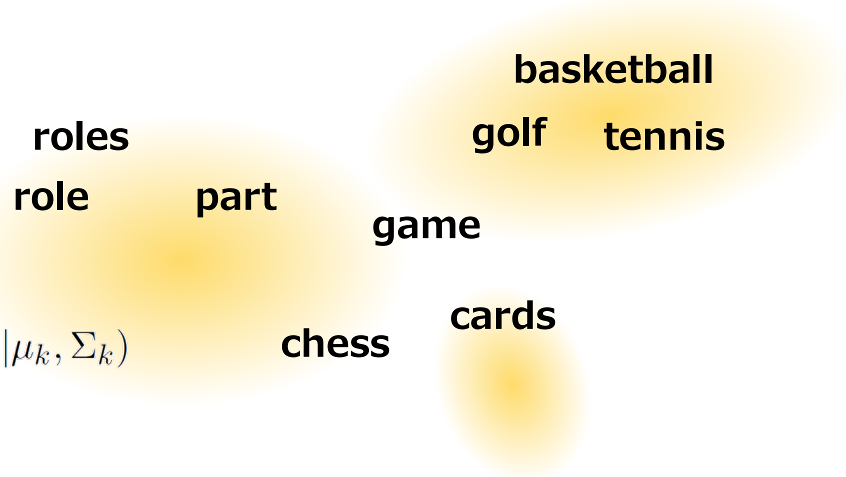
$$f_{\mathrm{CENT}}(w, W_c) = \cos(v_w, \frac{1}{|W_c|} \sum_{w_c \in W_c} v_{w_c})$$

(2) Gaussian model (GM)



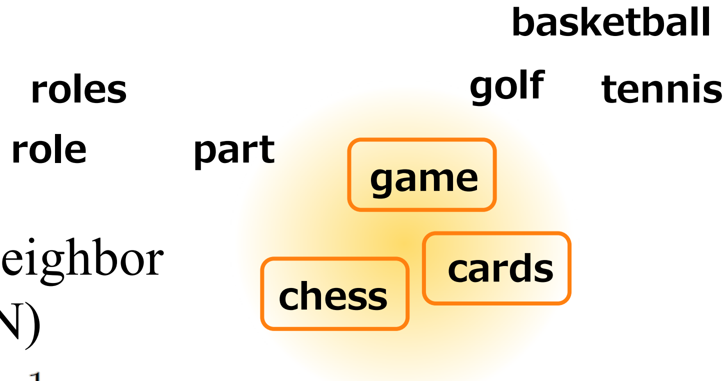
$$f_{\mathrm{GM}}(w, W_c) = \mathcal{N}(v_w | \mu, \Sigma)$$

(3) Gaussian mixture model (GMM)



$$f_{\mathrm{GMM}}(w, W_c) = \sum_{k=1}^{K} \pi_k \mathcal{N}(v_w | \mu_k, \Sigma_k)$$

(4) k-nearest neighbor model (kNN)



$$f_{k\mathrm{NN}}(w, W_c) = \frac{1}{k} \sum_{w_c \in k\mathrm{NN}_w(W_c)} \cos(v_w, v_{w_c})$$

(5) One-class SVM (1-SVM)

3 Models with negative instances

- Negative instances W_n: subset of words that *do not* belong to c & $W_n \cap W_o = \phi$

(6) OffSet-based model (OffSet)

$$f_{\mathrm{OffSet}}(w, W_c, W_n) = \cos\left(v_w, \frac{v_{\Sigma c}}{|v_{\Sigma c}|} - \frac{v_{\Sigma n}}{|v_{\Sigma n}|}\right)$$

$$\text{where } v_{\Sigma c} = \sum_{w_c \in W_c} v_{w_c}, \quad v_{\Sigma n} = \sum_{w_n \in W_n} v_{w_n}$$

(7) SVM with linear kernel (SVM_L)

(8) SVM with RBF kernel (SVM_R)

Experiments

Word embeddings & datasets

- 3 models (CBOW, SGNS, GloVe) for 2 languages
  - Use publicly available pre-trained word vectors for English
  - Train 300D embeddings on 1.5B word corpus for Japanese
- Selectional preference (SP) dataset
  - Sets of words that can be a direct object of a certain verb
  - e.g., {*role, part, game, golf, tennis*, etc.}
- WordNet dataset
  - Word sets extracted from English and Japanese WordNet
  - e.g., {*dog, llama, hedgehog, wolf*, etc.}

Experimental settings

- For each word set,
  - W_o is made by extracting 999 words from the other word sets
  - # of words for scoring is 1,000, including the target word w_t
  - W_n is also made similarly under the constraint $W_n \cap W_o = \phi$
  - Use 200 positive and 2,000 negative instances (i.e., $|W_c|$=200, $|W_n|$=2j,000)

- We regard the problem as a ranking task and adopt the mean reciprocal rank (MRR) as the metric for evaluation

$$\text{MRR} = \frac{1}{N}\sum_i^N \frac{1}{\text{rank}(w_{t_i})}$$

Results on the English SP dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1642*	.2539	.2360	.2097	.1726	.2782	.3397	**.3905**
SGNS	*.1887*	.2461	.2308	*.1918*	.2252	.2189	.3365	**.3608**
GloVe	*.1925*	.2596	.2462	.2245	.2295	.1150	.3554	**.3800**
Ave.	*.1818*	.2532	.2377	.2087	.2091	.2040	.3439	**.3771**

Results on the Japanese SP dataset.

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.2600*	.3151	.2947	.2783	.2812	.2516	.4371	**.4922**
SGNS	*.0789*	.2231	.2039	.1757	.1249	.2594	.4173	**.4510**
GloVe	*.1643*	.2489	.2377	.2016	.1927	.2088	.3264	**.3632**
Ave.	*.1677*	.2624	.2454	.2185	.1996	.2399	.3936	**.4355**

Results on the English SP dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1642*	.2539	.2360	.2097	.1726	.2782	.3397	**.3905**
SGNS	*.1887*	.2461	.2308	*.1918*	.2252	.2189	.3365	**.3608**
GloVe	*.1925*	.2596	.2462	.2245	.2295	.1150	.3554	**.3800**
Ave.	*.1818*	.2532	.2377	.2087	.2091	.2040	.3439	**.3771**

Results on the Japanese SP dataset.

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.2600*	.3151	.2947	.2783	.2812	.2516	.4371	**.4922**
SGNS	*.0789*	.2231	.2039	.1757	.1249	.2594	.4173	**.4510**
GloVe	*.1643*	.2489	.2377	.2016	.1927	.2088	.3264	**.3632**
Ave.	*.1677*	.2624	.2454	.2185	.1996	.2399	.3936	**.4355**

Results on the English SP dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1642*	.2539	.2360	.2097	.1726	.2782	.3397	**.3905**
SGNS	*.1887*	.2461	.2308	*.1918*	.2252	.2189	.3365	**.3608**
GloVe	*.1925*	.2596	.2462	.2245	.2295	.1150	.3554	**.3800**
Ave.	*.1818*	.2532	.2377	.2087	.2091	.2040	.3439	**.3771**

Results on the Japanese SP dataset.

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.2600*	.3151	.2947	.2783	.2812	.2516	.4371	**.4922**
SGNS	*.0789*	.2231	.2039	.1757	.1249	.2594	.4173	**.4510**
GloVe	*.1643*	.2489	.2377	.2016	.1927	.2088	.3264	**.3632**
Ave.	*.1677*	.2624	.2454	.2185	.1996	.2399	.3936	**.4355**

Results on the English WordNet dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1435*	*.1320*	*.1460*	*.1473*	*.1541*	.2263	.2564	**.2678**
SGNS	*.1767*	*.1679*	.1573	.1625	*.1704*	.1998	**.2292**	**.2357**
GloVe	*.1792*	*.1694*	.1562	*.1744*	*.1684*	.1310	.2075	**.2264**
Ave.	*.1665*	.1564	.1532	.1614	.1643	.1857	.2310	**.2433**

Results on the Japanese WordNet dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1996*	*.1991*	*.1918*	.2169	*.2082*	.2656	.2730	**.2961**
SGNS	*.0466*	.0521	.0774	.0768	.0701	.2367	.2686	**.2862**
GloVe	*.1055*	*.1050*	*.1021*	*.0987*	*.0984*	.0681	.2033	**.2189**
Ave.	*.1172*	.1187	.1238	.1308	.1256	.1901	.2483	**.2671**

Degree of membership

- Rosch developed the prototype concept and proved that not all members of a category are equally representative of the category

- Investigate how consistent the score calculated by each model is with Rosch's data on the degree of membership [Rosch'75]
  - College students are asked to use a 7-point scale to rate the extent to which each instance represents their idea or image of the category
  - We used eight categories that have a corresponding synset in WordNet

    e.g., Furniture: *chair*=1.04, *sofa*=1.04, *table*=1.1, ⋯, *stove*=5.4, ⋯

- Evaluate with Spearman's rank correlation coefficient (ρ) and Kendall's rank correlation coefficient (τ)

Averaged rank correlation coefficients

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
				ρ				
CBOW	.1736	.1905	.1706	.2417	.1160	**.3224**	.3176	.2562
SGNS	.2848	.3194	**.4024**	.3221	.1924	.2940	.3363	.3121
GloVe	.1458	.1949	.1448	.3204	.1780	**.4383**	.3367	.2702
Ave.	.2014	.2349	.2393	.2947	.1621	**.3516**	.3302	.2795
				τ				
CBOW	.1230	.1373	.1198	.1833	.0728	**.2400**	.2289	.1855
SGNS	.2101	.2400	**.2945**	.2355	.1400	.2066	.2390	.2180
GloVe	.1012	.1401	.1080	.2254	.1266	**.3038**	.2391	.1908
Ave.	.1448	.1725	.1741	.2147	.1131	**.2501**	.2357	.1981

Results on the English WordNet dataset

Model	CENT	GM	GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
CBOW	*.1435*	*.1320*	*.1460*	*.1473*	*.1541*	.2263	.2564	**.2678**
SGNS	*.1767*	*.1679*	.1573	.1625	*.1704*	.1998	**.2292**	**.2357**
GloVe	*.1792*	*.1694*	.1562	*.1744*	*.1684*	.1310	.2075	**.2264**
Ave.	*.1665*	.1564	.1532	.1614	.1643	.1857	.2310	**.2433**

correlation coefficients

			GMM	kNN	1-SVM	OffSet	SVM$_L$	SVM$_R$
				ρ				
CBOW	.1736	.1905	.1706	.2417	.1160	**.3224**	.3176	.2562
SGNS	.2848	.3194	**.4024**	.3221	.1924	.2940	.3363	.3121
GloVe	.1458	.1949	.1448	.3204	.1780	**.4383**	.3367	.2702
Ave.	.2014	.2349	.2393	.2947	.1621	**.3516**	.3302	.2795
				τ				
CBOW	.1230	.1373	.1198	.1833	.0728	**.2400**	.2289	.1855
SGNS	.2101	.2400	**.2945**	.2355	.1400	.2066	.2390	.2180
GloVe	.1012	.1401	.1080	.2254	.1266	**.3038**	.2391	.1908
Ave.	.1448	.1725	.1741	.2147	.1131	**.2501**	.2357	.1981

Conclusion

- Centroid-based approach cannot provide a reasonably good model
- Considering the geometry of the distribution and the existence of subgroups is useful but the impact is limited
- Negative instances must be taken into account for adequate modeling
- Discriminative learning-based models are best in finding the boundaries
- Offset-based models are best in determining the degree of membership