Talk@NCCU-CS, November 30, 2020

Network Embedding with Textual Information

Dr. Chuan-Ju Wang (王釧茹)
Associate Research Fellow
Research Center for Information Technology Innovation, Academia Sinica

Computational Finance and Data Analytics Laboratory (CFDA Lab)
http://cfda.csie.org

Outline

- Network Embedding with Textual Information
  - Item concept modeling
  - User review modeling
  - SIGIR'17, AAAI'19, TKDE'20

ICE: Item Concept Embedding via Textual Information

The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17), Tokyo, 2017, pp. 85-94. (full paper, acceptance rate: 22%)

https://dl.acm.org/citation.cfm?id=3080807

Extended version: **Item Concept Network: Towards Concept-based Item Representation Learning**, to appear in IEEE TKDE.

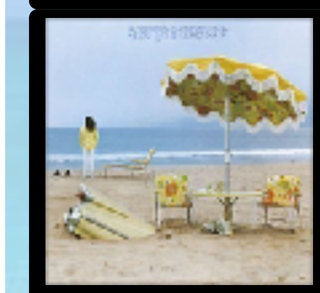
Normal search only retrieve the concept "beach"

beach

beach songs

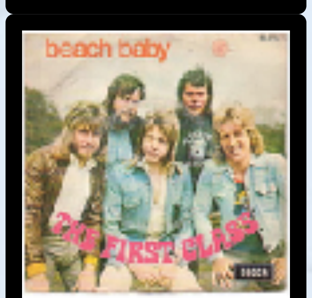
Girls on the **Beach**
Album: All Summer Long(1964)
Artist: The **Beach** Boys
… On the **beach** you'll find them there …

Rockaway **Beach**
Album: Rocket to Russia (1977)
Artist: Ramones
… Rock-rock, Rockaway **Beach** …



On the **Beach**
Album: On the **Beach** (1974)
Artist: Neil Young
… out here on the **beach** …

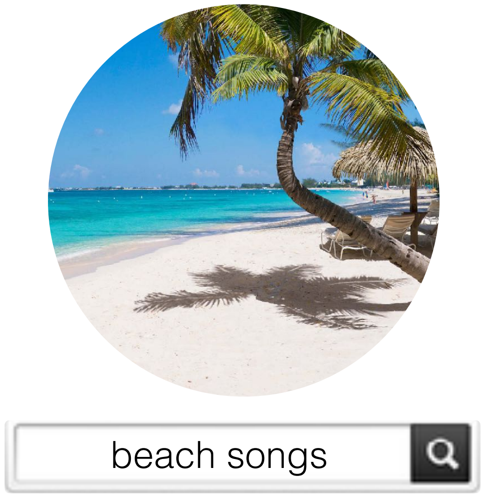
Private **Beach** Party
Album: Private **Beach** Party (1985)
Artist: Gregory Isaacs
… At the private **beach** party…



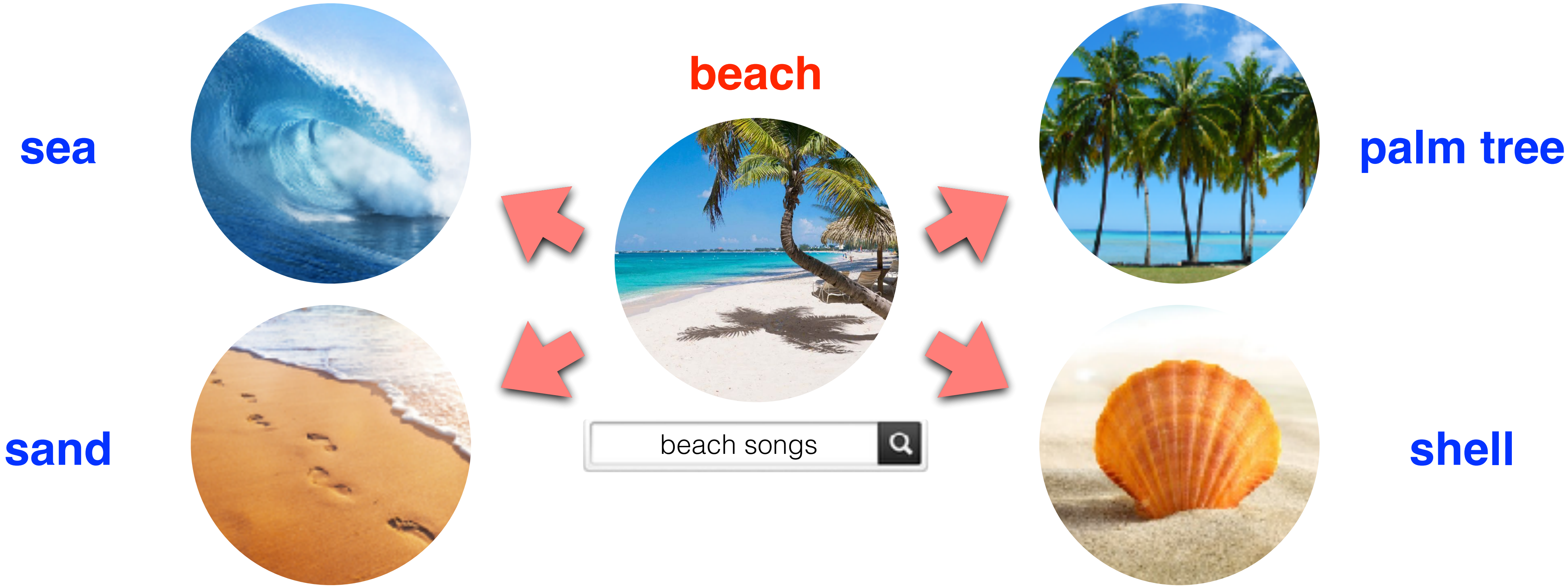
Private **Beach** Baby
Album: single (1974)
Artist: The First Class
… **Beach** baby, **beach** baby…

“Beach” has many correlated concepts

beach



Expand concept "beach" to sea, sand, …



Capture similar concepts = diverse AND relevant

beach

sea

palm tree

sand

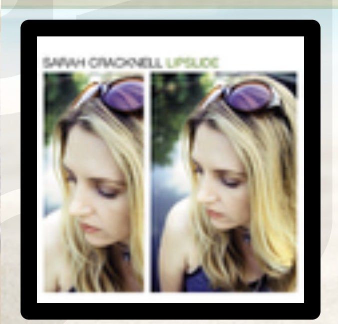
shell

beach songs

Girls on the **Beach**
Album: All Summer Long (1964)
Artist: The **Beach** Boys
… On the **beach** you'll find them

Palmtree
Album: single (2015)
Artist: Mandelbarth
… Under the **palm trees** is where we …

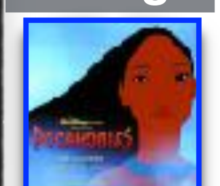
Sand And **Sea**
Album: That's Life (1966)
Artist: Frank Sinatra
… **Sand** and **sea**, **sea** and **sand** …



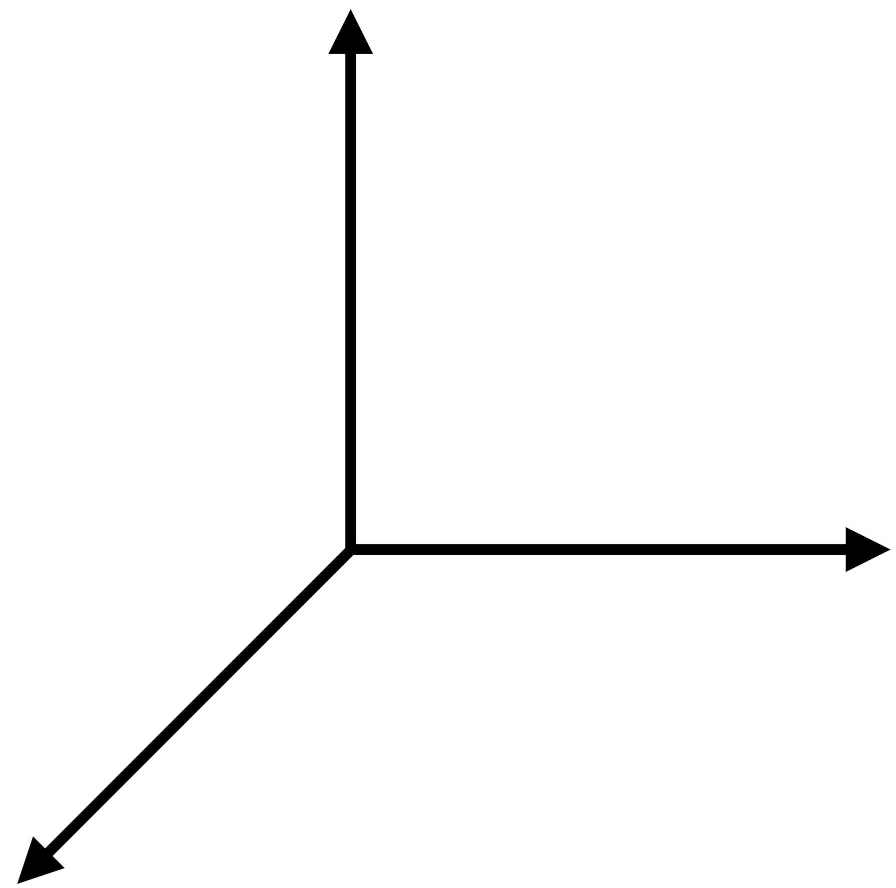
Sea **Shells**
Album: Lipslide (1997)
Artist: Sarah Cracknell
… Hey little **sea shell**, I need a cue…

Embed items and concepts in space such that…

Song	Lyrics	Word
	… On the **beach** you'll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**
	… Hey little **sea shell**, I need a cue…	**sea** **shell**
	… Under the **palm trees** is where we …	**palm tree**

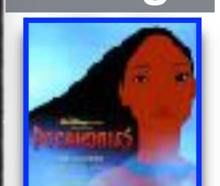
Song	Lyrics	Word
	… all the voices of the **mountains** … All you own is **earth** until …	**mountain** **earth**
	… Far away o'er the **mountains**, … with the **cliffs** of Doneen …	**mountain** **cliff**

Embedding Space

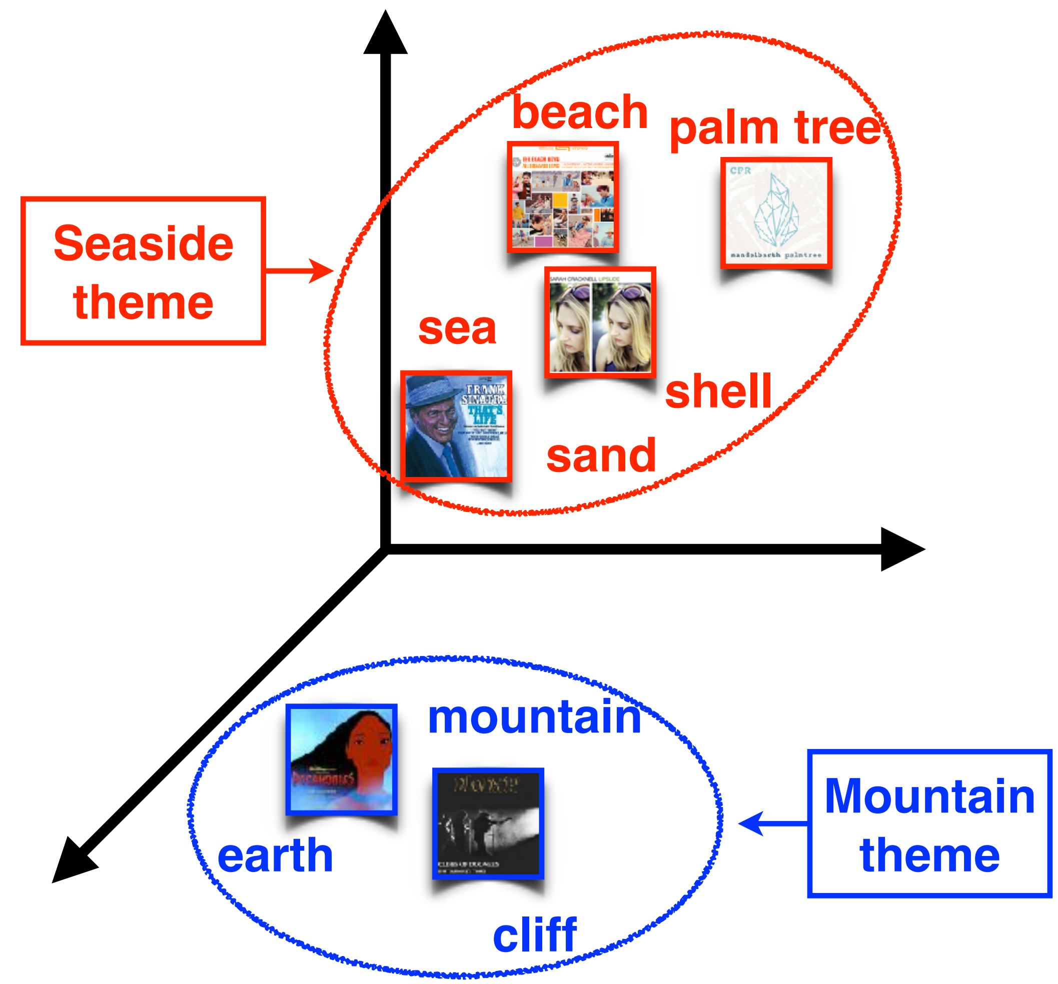


… similar items and concepts flock together

Song	Lyrics	Word
	… On the **beach** you'll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**
	… Hey little **sea shell**, I need a cue…	**sea** **shell**
	… Under the **palm trees** is where we …	**palm tree**

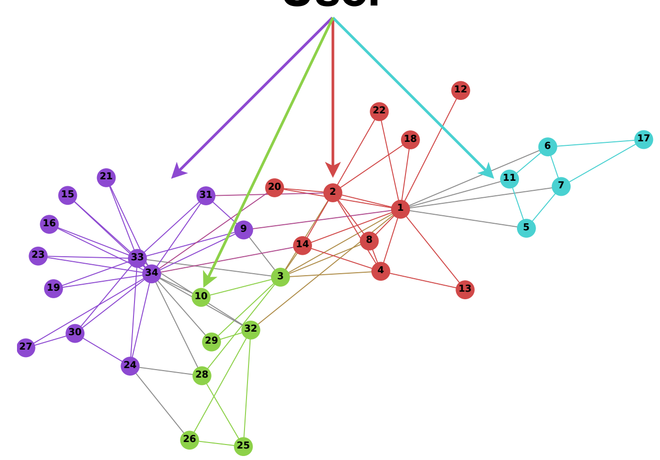
Song	Lyrics	Word
	… all the voices of the **mountains** … All you own is **earth** until …	**mountain** **earth**
	… Far away o'er the **mountains**, … with the **cliffs** of Doneen …	**mountain** **cliff**

Embedding Space

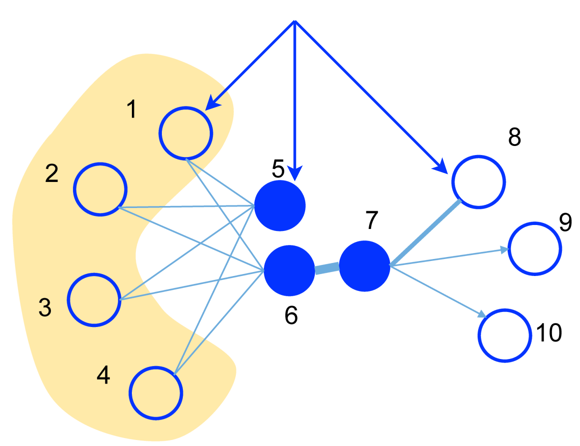


… and different ones separate.

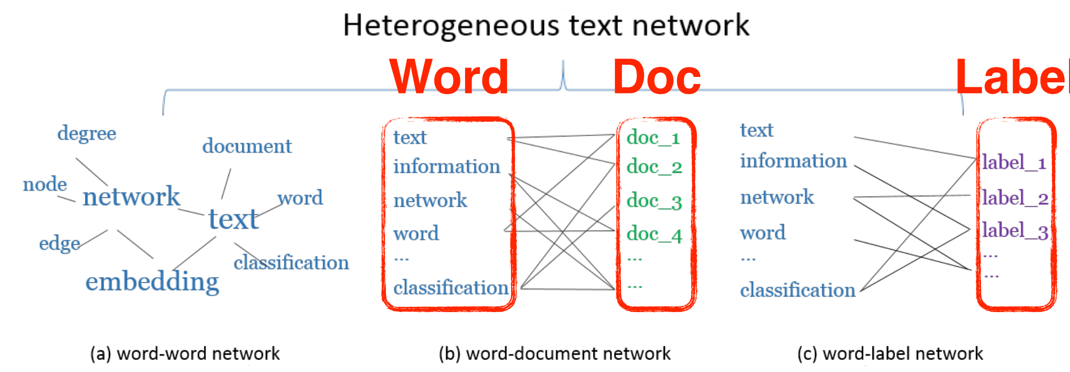
Related works in graph embedding



DeepWalk (Perozzi et al., 2014)



LINE (Tang, et al., 2015)



PTE (Tang, et al. 2015)

- All the above-mentioned methods focus on **homogeneous** tasks:
  - **DeepWalk**: Homogeneous social networks (users with social relations).
  - **LINE**: Homogeneous social networks or word-word networks, etc.
  - **PTE**: Heterogeneous text network but still for homogeneous tasks, such as document classification.
- However, the inter-retrieval task between concepts and items is **heterogeneous**:
  - e.g., word-to-song retrieval, movie-to-word retrieval, etc.

Our Proposal: Item Concept Embedding (ICE)

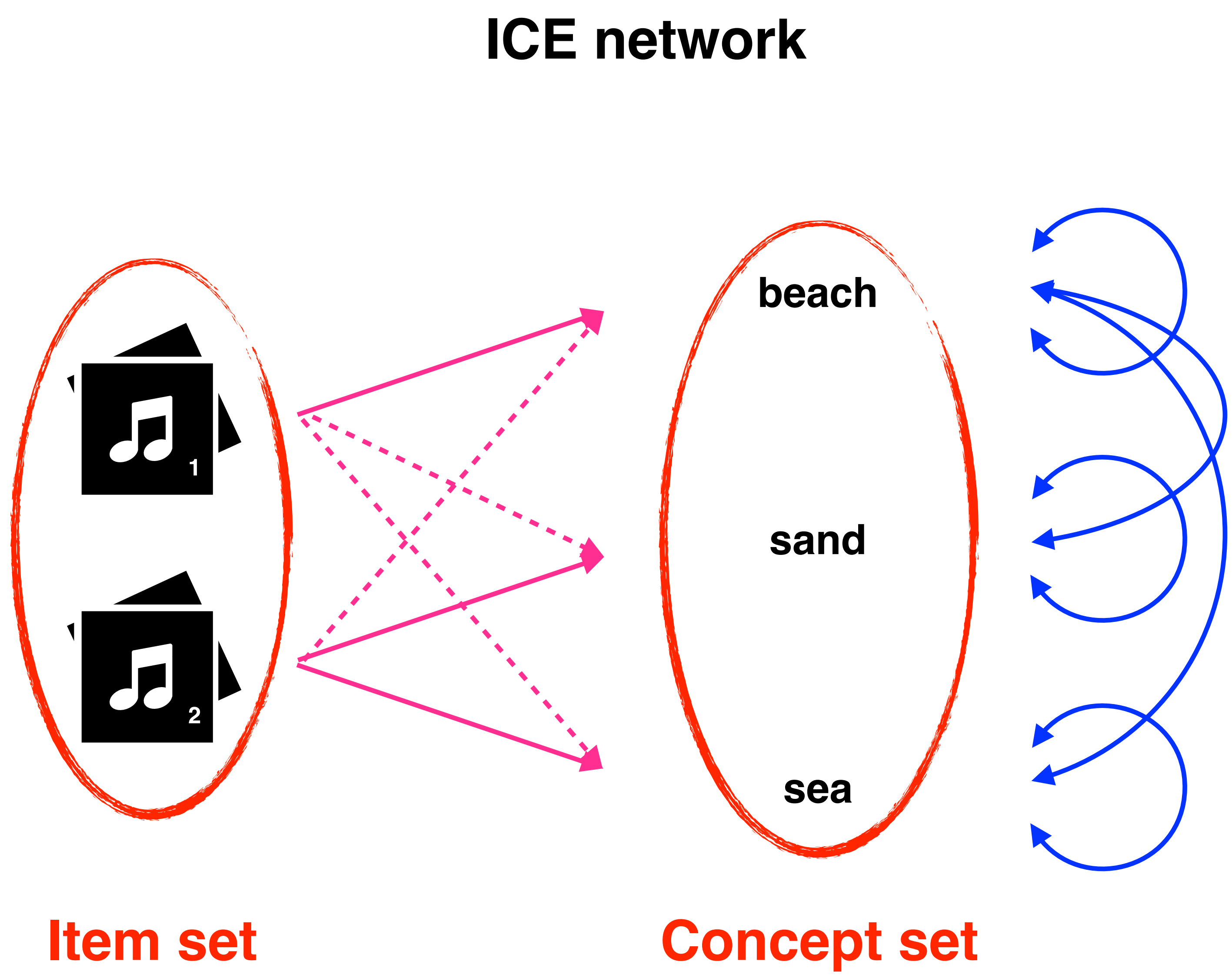
- Main Contributions:

  1. Propose item concept embedding (ICE) approach to model the concepts of items via associated textual information.

  2. Integrate heterogeneous nodes and relations in network using generalized matrix operations.

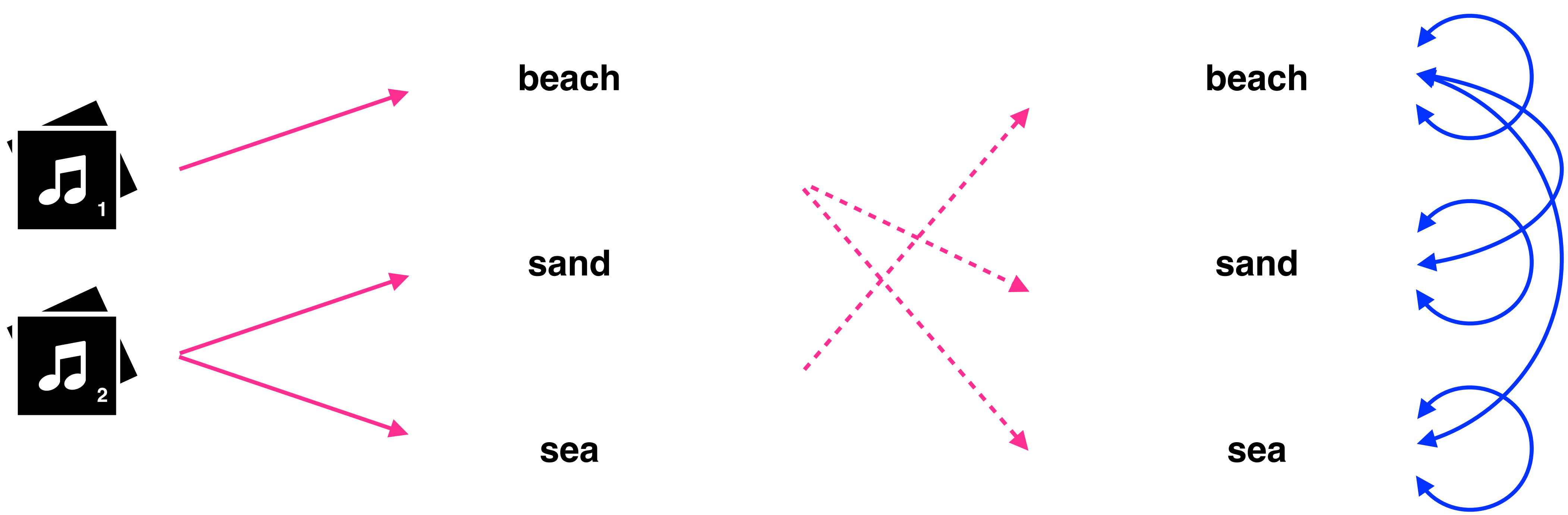
  3. Learn embeddings capable to retrieve conceptually diverse and relevant results that support both homogeneous and heterogeneous tasks.

ICE network is an unified network composed of…



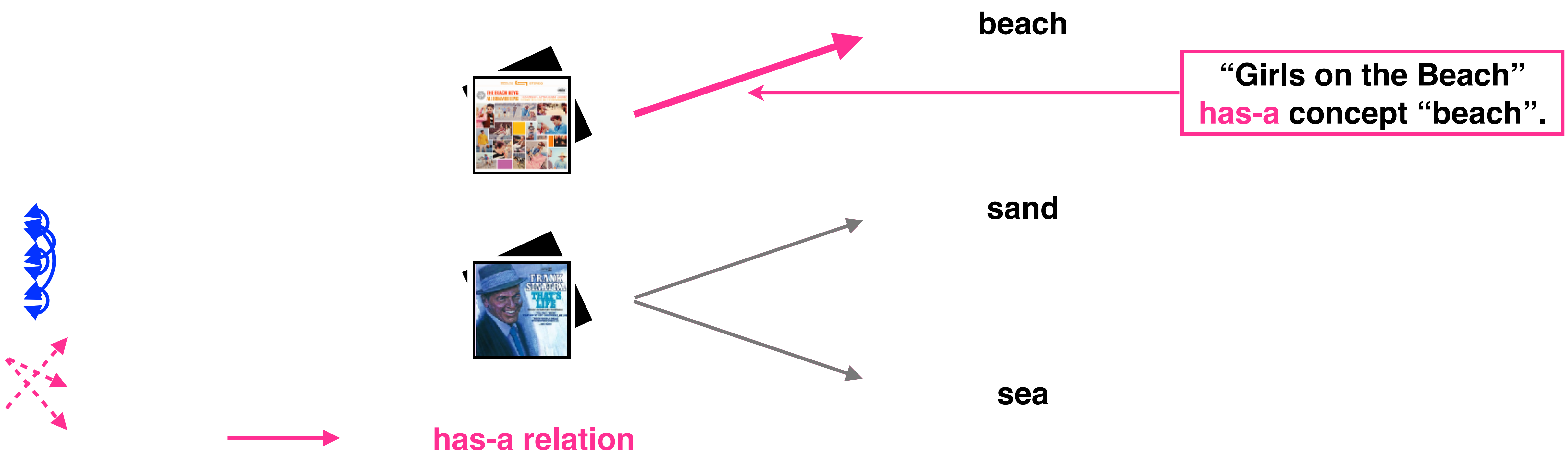
… 2 basis networks and 3 relations

Entity-text network **Text-text network**



Has-a relation Expanded has-a relation Concept-similar relation

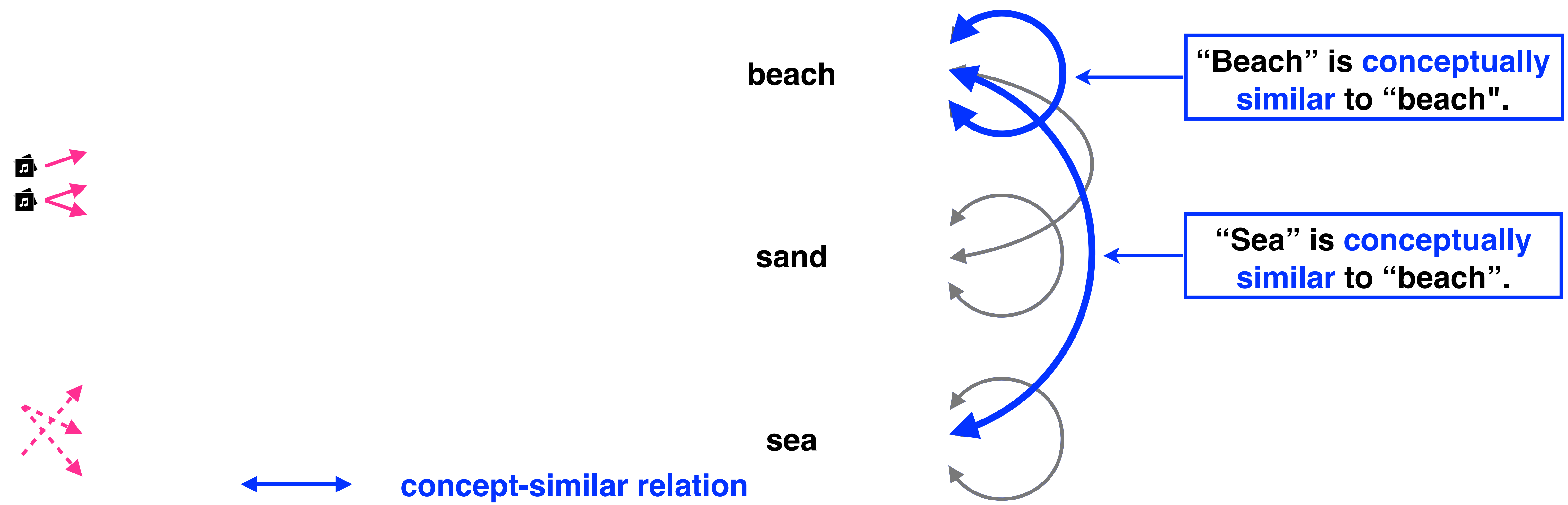
Entity-text network manages item concepts



Songs	Lyrics	Words
	… On the **beach** you'll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**

- Manage the has-a relation between each item and their representative concept words.
- Concept words for each item are picked according to the TF-IDF score.
- Heterogeneous, directed, and bipartite.

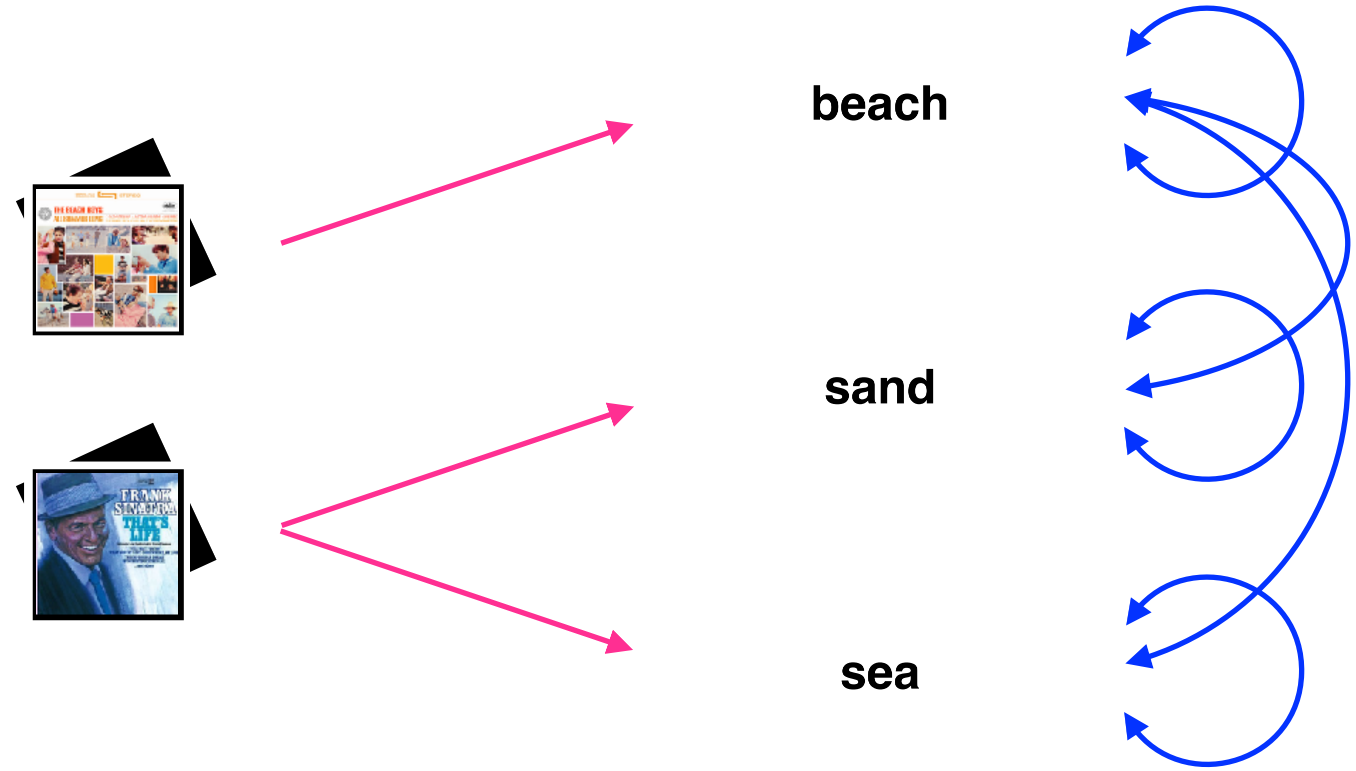
Text-text network manages concept similarity



Songs	Lyrics	Words
	… On the **beach** you’ll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**

- Manage the concept-similar relation between each concept words.
- Conceptually similar words are connected according to the cosine similarity between word embeddings.
- Homogeneous and bi-directed.

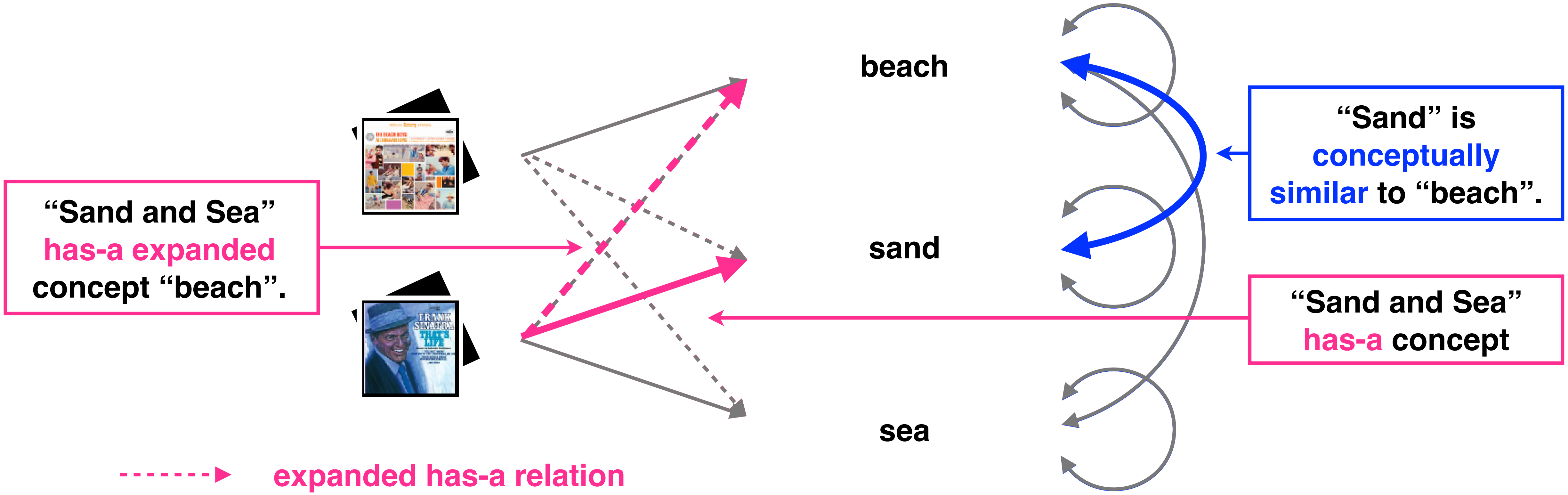
ICE network combines E-T and T-T network and …



Songs	Lyrics	Words
	… On the **beach** you'll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**

- Combine entity-text network, text-text network, and expanded has-a relation.
- Manage the expanded has-a relation between each item and their expanded concept words.
- Establish relation to expanded concept words via the conceptually similar words of each item.
- Heterogeneous nodes and relations.

… manages the expanded has-a relation

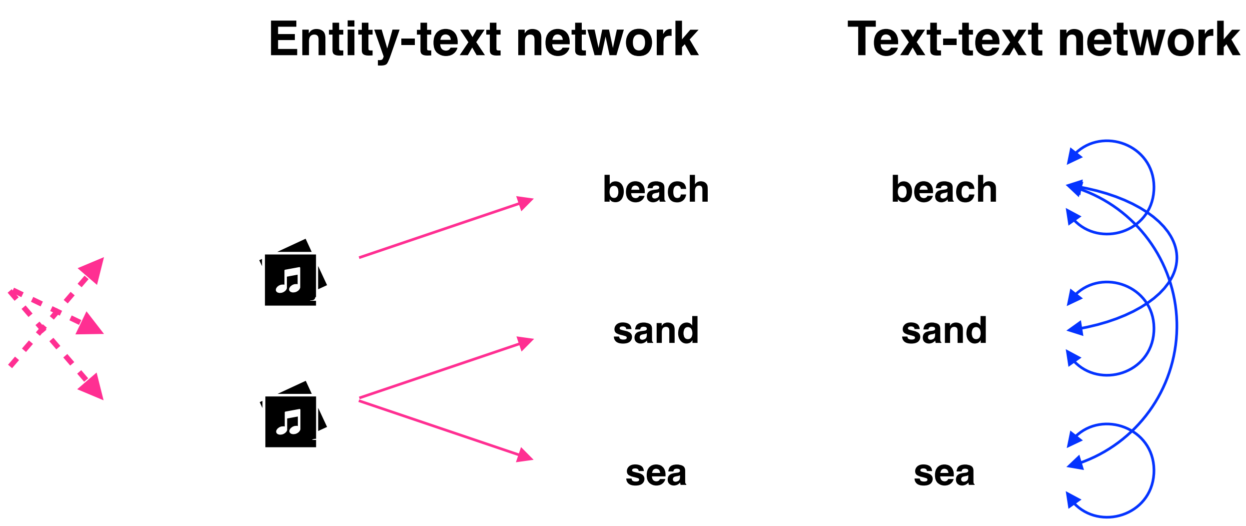


Songs	Lyrics	Words
	… On the **beach** you'll find them there …	**beach**
	… **Sand** and **sea**, **sea** and **sand** …	**sand** **sea**

- Combine entity-text network, text-text network, and expanded has-a relation.
- Manage the expanded has-a relation between each item and their expanded concept words.
- Establish relation to expanded concept words via the conceptually similar words of each item.
- Heterogeneous nodes and relations.

Construct graph via generalized matrix operation

- Step 1: Establish expanded has-a relation in ET network.



Construct graph via generalized matrix operation

- Step 1: Establish expanded has-a relation in ET network.

Entity-text network

$$\begin{array}{c|ccc} & W_1 & W_2 & W_3 \\ \hline I_1 & 1 & 0 & 0 \\ I_2 & 0 & 1 & 1 \end{array}$$

$$M_{G_{et}}$$

Text-text network

$$\begin{array}{c|ccc} & W_1 & W_2 & W_3 \\ \hline W_1 & 1 & 1 & 1 \\ W_2 & 1 & 1 & 0 \\ W_3 & 1 & 0 & 1 \end{array}$$

$$M_{G_{tt}}$$

Construct graph via generalized matrix operation

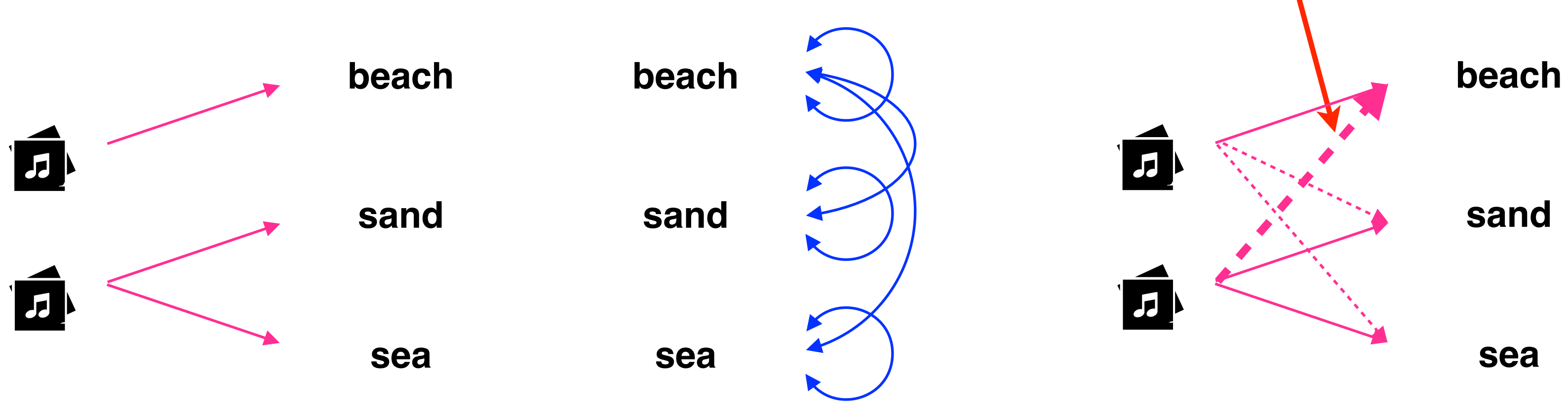
- Step 1: Establish expanded has-a relation in ET network.

Entity-text network **Text-text network** $A = M_{G_{et}} \cdot M_{G_{tt}}$

$$\begin{array}{c} \\ I_1 \\ I_2 \end{array}\overset{\begin{array}{ccc} W_1 & W_2 & W_3 \end{array}}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}} \cdot \begin{array}{c} \\ W_1 \\ W_2 \\ W_3 \end{array}\overset{\begin{array}{ccc} W_1 & W_2 & W_3 \end{array}}{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}} = \begin{array}{c} \\ I_1 \\ I_2 \end{array}\overset{\begin{array}{ccc} W_1 & W_2 & W_3 \end{array}}{\begin{bmatrix} 1 & 1 & 1 \\ \boxed{2} & 1 & 1 \end{bmatrix}}$$

$M_{G_{et}}$ $M_{G_{tt}}$

More encompassing concept



Construct graph via generalized matrix operation

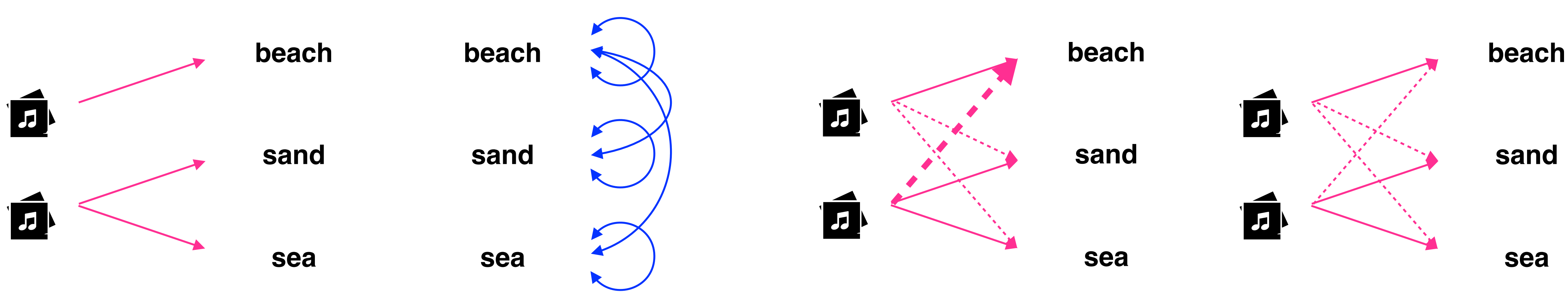
- Step 2: Convert the dot product to a binary matrix $\tilde{A}$.

Entity-text network **Text-text network**

$$A = M_{G_{et}} \cdot M_{G_{tt}} \qquad \tilde{A} = \left(\mathbb{1}_{\{a_{ij}>0\}}\right)$$

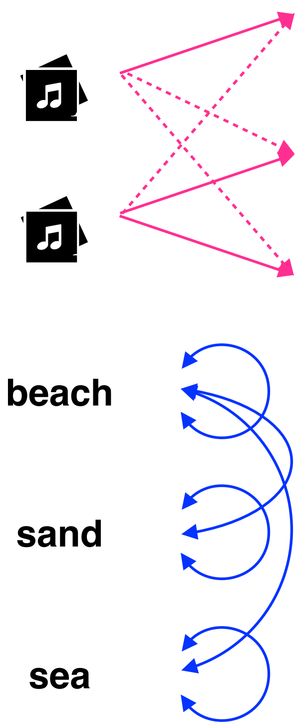
$$\begin{array}{c c} & \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{array}{c} I_1 \\ I_2 \end{array} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \cdot \begin{array}{c c} & \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{array}{c} W_1 \\ W_2 \\ W_3 \end{array} & \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \end{array} = \begin{array}{c c} & \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{array}{c} I_1 \\ I_2 \end{array} & \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c c} & \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{array}{c} I_1 \\ I_2 \end{array} & \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

$M_{G_{et}}$ $M_{G_{tt}}$



Construct graph via generalized matrix operation

- Step 3: Augment binary matrix with the text-text matrix.



$$\tilde{A}$$

$$\begin{array}{c} \\ I_1 \\ I_2 \end{array}\begin{array}{c} \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ W_1 \\ W_2 \\ W_3 \end{array}\begin{array}{c} \begin{array}{ccc} W_1 & W_2 & W_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$M_{G_{tt}}$$

Construct graph via generalized matrix operation

- Step 3: Augment binary matrix with the text-text matrix.

$$\tilde{A}$$

$$\begin{array}{c|c} & \begin{matrix} W_1 & W_2 & W_3 \end{matrix} \\ \begin{matrix} I_1 \\ I_2 \end{matrix} & \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

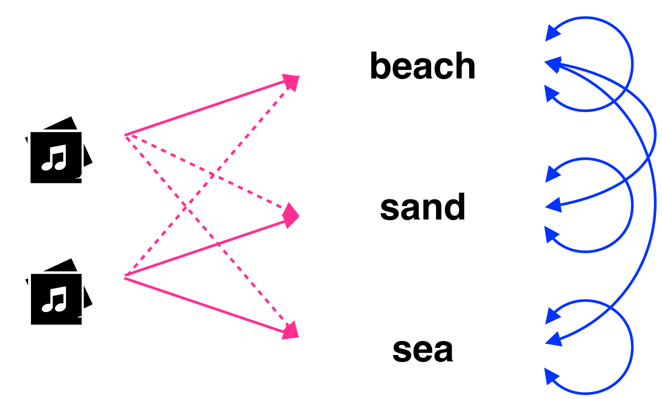
$$\begin{array}{c|c} & \begin{matrix} W_1 & W_2 & W_3 \end{matrix} \\ \begin{matrix} W_1 \\ W_2 \\ W_3 \end{matrix} & \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$M_{G_{tt}}$$

$$M_{G_{ice}} = \begin{bmatrix} \tilde{A} \\ M_{G_{tt}} \end{bmatrix}$$

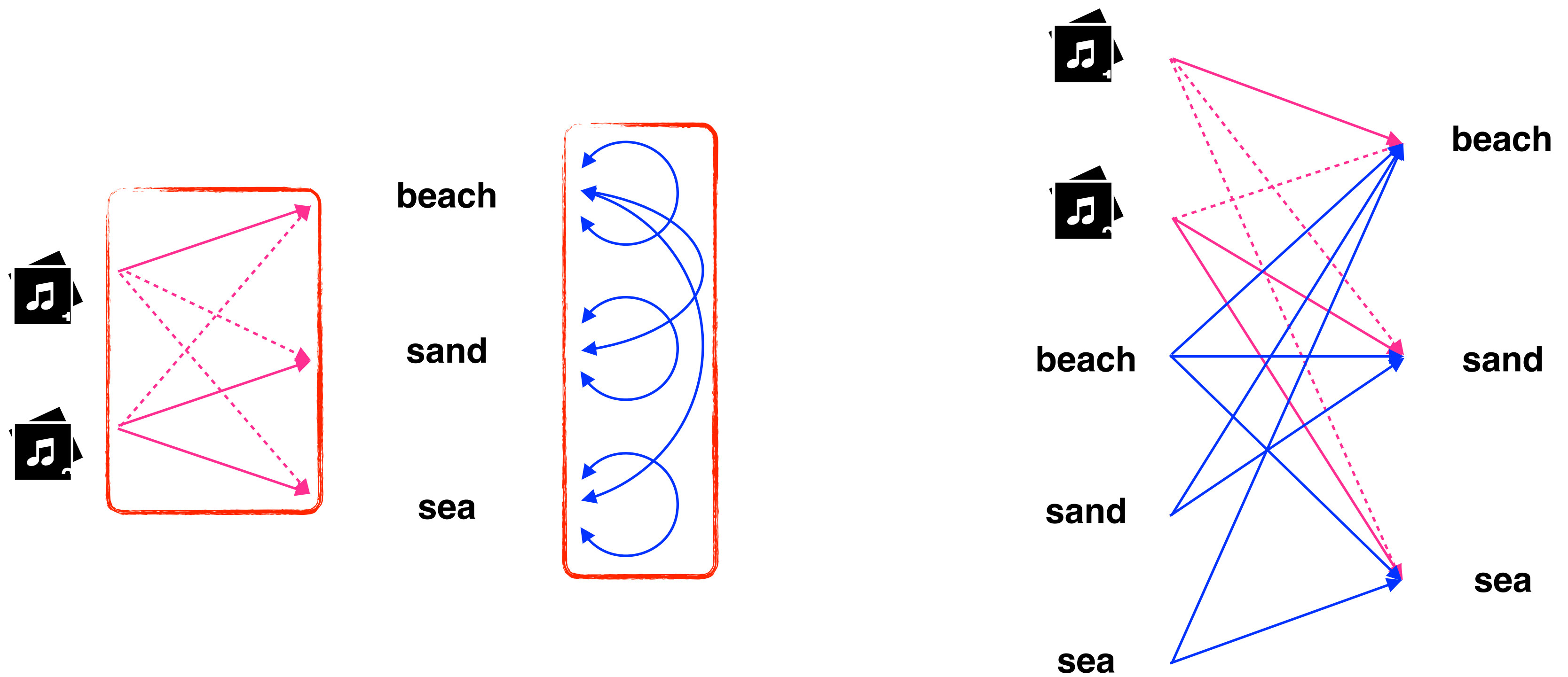
$$\begin{array}{c|c} & \begin{matrix} W_1 & W_2 & W_3 \end{matrix} \\ \begin{matrix} I_1 \\ I_2 \\ W_1 \\ W_2 \\ W_3 \end{matrix} & \left[\begin{array}{ccc} 1 & 1 & 1 \\ 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right] \end{array}$$

G_{ice}: ICE Network



Modeling of neighborhood proximity

- Intuition: Maintain homogeneous neighborhood.



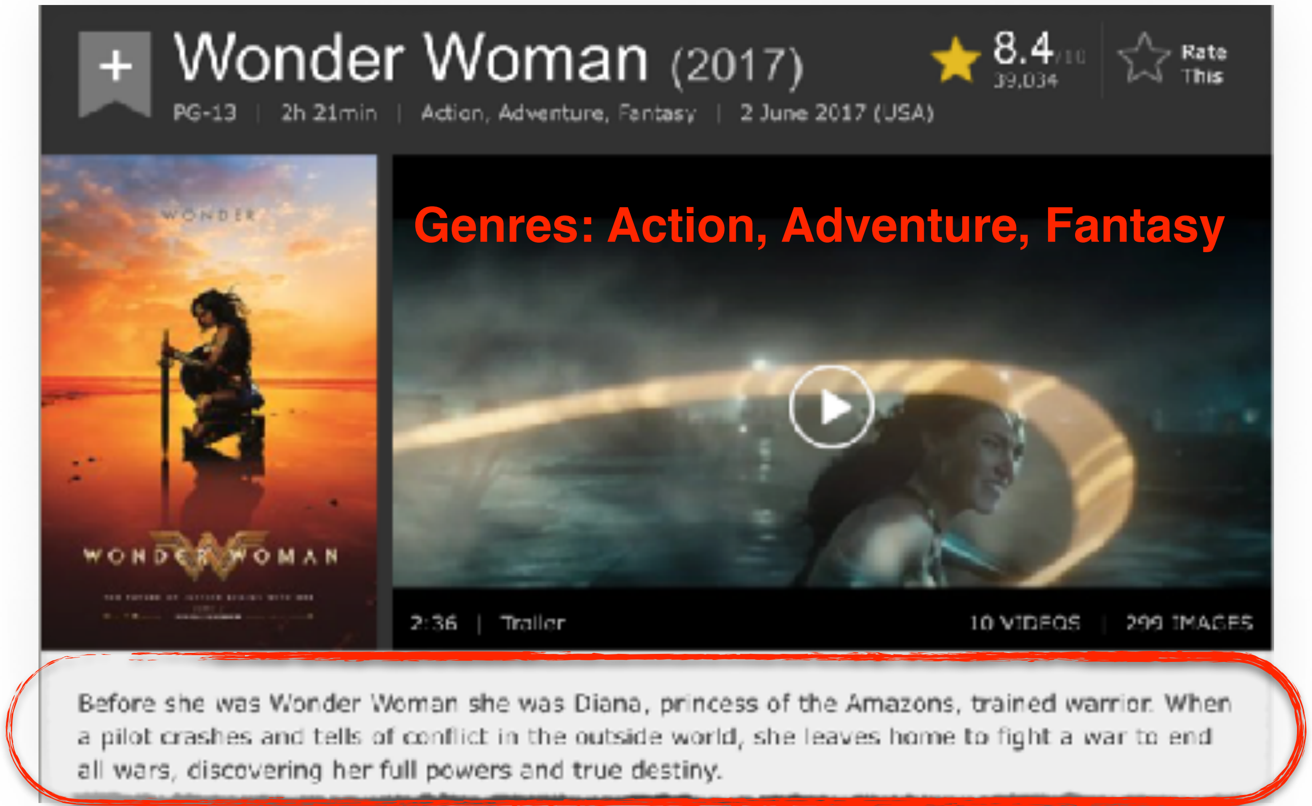
(a) ICE network (b) Homogeneous context

- Jointly minimize the KL divergence of objective functions:

$$O_{ice} = -\left(\sum_{(n_i, n_\ell) \in \tilde{E}_{et}} x_{i\ell} \log P(n_\ell \mid n_i) + \sum_{(n_w, n_\ell) \in E_{tt}} x_{w\ell} \log P(n_\ell \mid n_w) \right)$$

Datasets: Real-world movie and music datasets

- IMDB (movie) dataset:
  - Movie, plots, and genres
- KKBOX (music) dataset:
  - Song and lyrics



	IMDB	KKBOX
# movies/songs	36,586	33,106
Average text length	65.0	215.24
Average # unique words	47.8	81.37
Vocabulary size	66,924	101,395
# single genres	28	-
# multi-label genres	915	-

稻香 — 周杰倫（Jay Chou）

對 / 這個 / 世界 / 如果/ 你 / 有 / 太多 / 的 / 抱怨
跌倒 / 了 / 就 / 不敢 / 繼續/ 往前 / 走
為什麼 / 人 / 要 / 這麼 / 的 / 脆弱/ / 墮落
請 / 你 / 打開 / 電視 / 看看
多少 / 人 / 為 / 生命 / 在 / 努力 / 勇敢 / 的 / 走 / 下去
我們 / 是不是 / 該 / 知足
珍惜 / 一切 / 就算 / 沒有 / 擁有
…

Experiment — Tasks and baselines

- Two types of tasks:
  1. Homogeneous:
     - Movie classification.
     - Movie-to-movie retrieval.
  2. Heterogeneous:
     - Word-to-movie retrieval. (Ex: Using "Killer" in Thriller movies.)
     - Movie-to-word retrieval.
     - Word-to-song retrieval. (Ex: Using contextual words.)
- Baselines:
  1. Traditional: Keyword-based (**KBR**), bag-of-words (**BOW**)
  2. Embedding: Bipartite (**BPT**), average embedding (**AVGEMB**)

Homogeneous: Movie genre classification

- Multi-label Movie Genre Classification (homogeneous):

Table 4: Movie genre classification task

W = # of rep words per item

exp = # of exp. words per concept word

	$\|W\| = 10$					$\|W\| = 20$			
	BOW	BPT	ICE (exp-3)	ICE (exp-5)		BOW	BPT	ICE (exp-3)	ICE (exp-5)
Exact match ratio	0.136	**0.160**	0.156	0.157		0.162	**0.182**	**0.182**	0.181
Micro-average F-measure	0.365	0.401	0.408	**0.410**	<	0.415	**0.464**	0.462	0.463
Macro-average F-measure	0.087	0.166	**0.170**	**0.170**		0.156	**0.229**	0.223	0.222

- Increasing the number of concept words used to represent an item improves the performance of the item embedding.

Comparable performance in homogeneous tasks

- Multi-label Movie Genre Classification (homogeneous):

Table 4: Movie genre classification task

W = # of concept words per item

exp = # of exp. words per concept word

	$\|W\| = 10$				$\|W\| = 20$			
	BOW	BPT	ICE (exp-3)	ICE (exp-5)	BOW	BPT	ICE (exp-3)	ICE (exp-5)
Exact match ratio	0.136	**0.160**	0.156	0.157	0.162	**0.182**	**0.182**	0.181
Micro-average F-measure	0.365	0.401 ≈	0.408 ≈	**0.410**	0.415	**0.464** ≈	0.462 ≈	0.463
Macro-average F-measure	0.087	0.166	**0.170**	**0.170**	0.156	**0.229**	0.223	0.222

- ICE embeddings are suitable for homogeneous tasks.

Heterogeneous: Word-to-movie retrieval

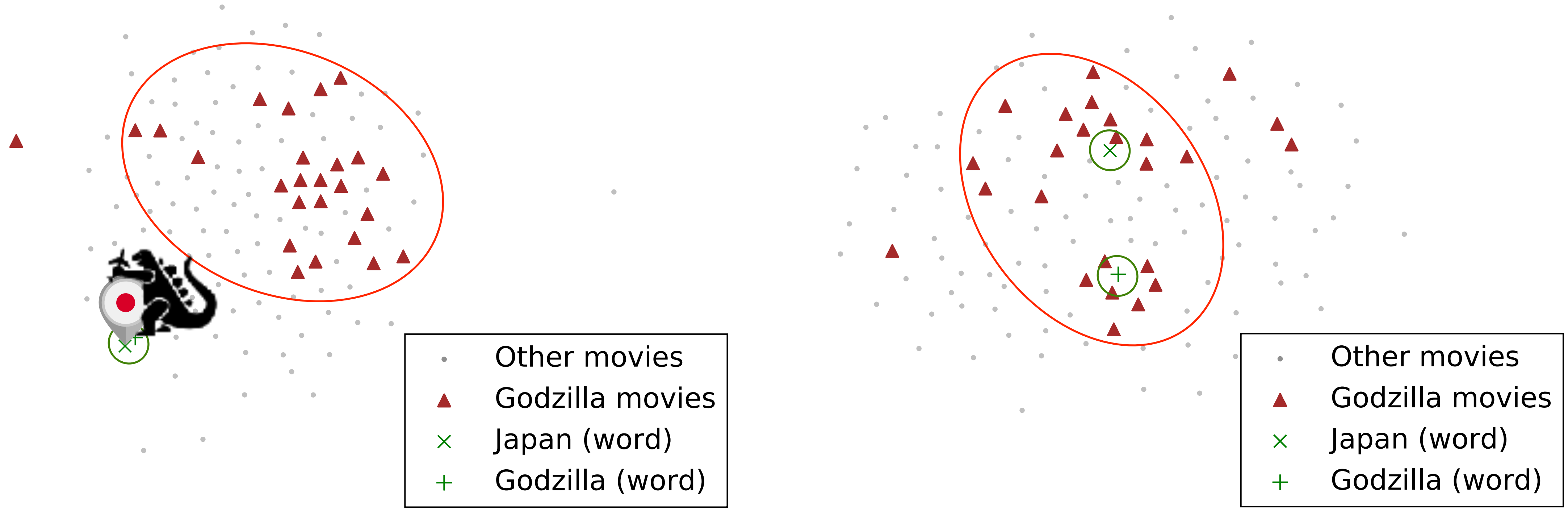
- Word-to-movie Retrieval (heterogeneous):

Table 5: Word-to-movie retrieval task

$\lvert W \rvert = 20$	Horror (3754/36586)	Thriller (4636/36586)	Western (751/36586)	Action (5029/36586)	Short (1094/36586)	Sci-Fi (2004/36586)	Average
		"Killer"		P@50		"Alien"	
RAND	0.080	0.080	0.060	0.080	0.000	0.120	0.070
KBR	0.324	**0.230**	**0.321**	0.418	0.062	0.373	0.288
AVGEMB	0.322	0.212	0.316	0.406	0.092	0.392	0.290
AVGEMB (all)	0.324	0.225	0.304	0.366	0.089	**0.401**	0.285
BPT	0.096	0.104	0.010	0.154	0.032	0.086	0.080
ICE (exp-5)	**0.354**	0.204	0.294	**0.444**	**0.142**	0.392	**0.305**
				P@100			
RAND	0.050	0.100	0.030	0.110	0.000	0.060	0.058
KBR	**0.327**	**0.224**	0.236	0.395	0.057	0.307	0.258
AVGEMB	0.324	0.215	0.266	0.385	0.074	0.372	0.273
AVGEMB (all)	0.314	0.208	**0.269**	0.376	0.074	**0.382**	0.270
BPT	0.088	0.116	0.012	0.156	0.034	0.086	0.082
ICE (exp-5)	0.321	0.193	0.264	**0.421**	**0.109**	0.362	**0.278**

Movies flock to concepts with high similarity

Figure 4: Visualization of the Representations of the Godzilla-related Movies and Two Related Keywords

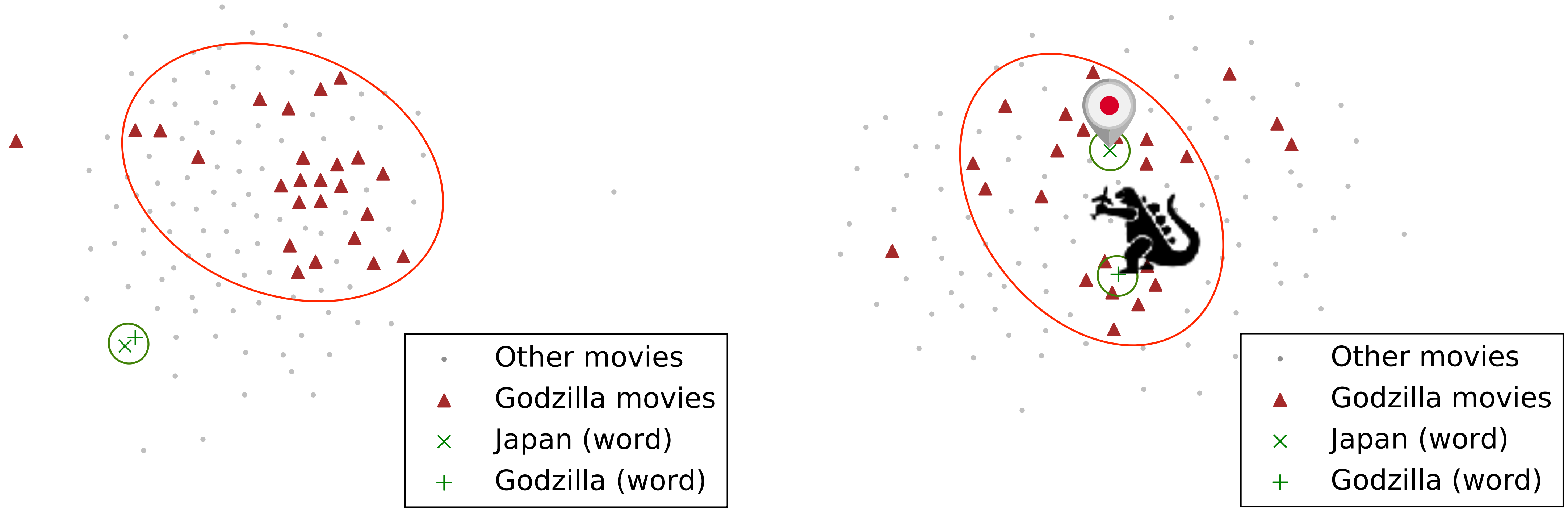


(a) BPT with $|W| = 20$

(b) ICE (exp-5) with $|W| = 20$

Movies flock to concepts with high similarity

Figure 4: Visualization of the Representations of the Godzilla-related Movies and Two Related Keywords



(a) BPT with $|W| = 20$

(b) ICE (exp-5) with $|W| = 20$

- ICE concept embeddings can retrieve movies of similar concepts, and vice versa.
- Therefore, ICE embeddings are suitable for heterogeneous tasks.

Heterogeneous: Word-to-song retrieval

- Word-to-song Retrieval (heterogeneous):

Table 6: Performance comparison on the 15 keywords

$\lvert W \rvert = 10$		Keyword				Concept-similar word			
			P@100				P@100		
	Query	# keyword songs	BPT	AVGEMB	ICE (exp-3)	# concept-similar songs	BPT	AVGEMB	ICE (exp-3)
Mood	失落 (lost)	516	0.000	0.160	**0.470**	403	0.030	**0.120**	0.050
	心痛 (heartache)	824	0.050	0.080	**0.250**	4,075	0.170	0.500	**0.610**
	想念 (pining)	1,729	0.050	0.250	**0.700**	1,176	0.080	**0.180**	0.060
	深爱 (affectionate)	380	0.000	0.090	**0.550**	442	0.020	0.110	**0.250**
	難過 (sad)	1678	0.040	0.200	**0.530**	1,781	0.080	**0.320**	0.070
Location	回家 (home)	934	0.040	0.310	**0.900**	1,190	0.020	**0.340**	0.160
	房間 (room)	610	0.000	0.420	**0.510**	28	0.000	0.010	**0.060**
	海邊 (seaside)	264	0.000	0.230	**0.360**	91	0.000	0.070	**0.080**
	火車 (train)	151	0.010	0.330	**0.510**	20	0.000	**0.040**	0.020
	花園 (garden)	139	0.000	0.160	**0.390**	2	0.000	0.000	0.000
Time	夕陽 (dusk)	387	0.010	0.180	**0.360**	307	0.020	**0.100**	0.070
	日出 (sunrise)	240	0.000	0.290	**0.430**	390	0.060	0.380	**0.690**
	日落 (sunset)	226	0.030	0.380	**0.590**	407	0.010	0.270	**0.530**
	月亮 (moon)	598	0.000	0.360	**0.930**	1,608	0.030	0.320	**0.350**
	黑夜 (dark night)	1,189	0.030	0.140	**0.510**	279	**0.030**	**0.030**	0.010
	Total/Avg. P@100	9,865	0.017	0.239	**0.533**	12,199	0.037	0.186	**0.201**

Context types

Diverse and relevant by ConceptNet

Table 7: Performance evaluated by ConceptNet Human-labeled semantic knowledge graph

$\|W\| = 10$		P@10		Diversity@10		P@100		Diversity@100	
Query	# words in ConceptNet	KBR	ICE (exp-3)	KBR	ICE (exp-3)	KBR	ICE (exp-3)	KBR	ICE (exp-3)
夕陽 (dusk)	11	0.00	**0.20**	0.00	0.00	**0.25**	0.08	0.00	**0.75**
房間 (room)	39	**0.60**	0.10	0.00	0.00	**0.36**	0.16	0.00	**0.69**
日出 (sunrise)	17	0.40	**1.00**	0.00	**0.70**	**0.30**	0.24	0.00	**0.75**
花園 (garden)	33	**0.30**	0.10	0.00	0.00	**0.34**	0.08	0.00	**0.50**
黑夜 (dark night)	17	0.50	**1.00**	0.00	0.00	0.50	**0.57**	0.00	**0.68**
Average	23.4	0.36	**0.48**	0.00	**0.14**	**0.35**	0.23	0.00	**0.67**

Relevance Diversity

- Songs retrieved using ICE word embeddings have high diversity and relevance by human standard.

Case Study

Table 8: An example for movie-to-word retrieval

Query movie: Toy Story, 1995 (Animation, Adventure, Comedy)		
BPT	ICE (exp-5)	
manias	andy	Protagonist
entraineuse	gave	
taddeo	give	
anuelo	sid	Antagonist
portico	tabbed	
bep	robertson	
meanness	Named	
zanchi	stuffed_animals	Generic toys
sarti	toys	
raffin	Toys	

Table 10: An example for word-to-movie retrieval

Word query: $\boxed{\text{alien}}$ **Representative concept for Sci-Fi**	
BPT	ICE (exp-5)
The Blue Lagoon, 1949 (Adventure, Drama, Romance)	Coneheads, 1993 (Comedy, **Sci-Fi**)
Turner & Hooch, 1989 (Comedy, Crime, Drama)	Without Warning, 1980 (**Sci-Fi**, Horror)
Only the Young, 2012 (Documentary, Comedy, Romance)	They Came from Beyond Space, 1967 (Adventure, **Sci-Fi**)
Brute Force, 1947 (Crime, Drama, Film-Noir)	Battle of the Stars, 1978 (**Sci-Fi**)
Home, 2015 (Animation, Adventure, Comedy)	Howard the Duck, 1986 (Action, Adventure, Comedy)

Short Recap

1. Propose the ICE framework, which models item concepts using textual information.
2. Propose a generalized network construction method based on matrix operations.
3. Leverage neighborhood proximity to learn embeddings capable to be used in both homogeneous and heterogeneous tasks.
4. Resulted embeddings can be used to retrieve conceptually diverse an relevant items.

Release: ICE API and dataset

git
hub.com/
cnclabs/

- ICE API:
  - Repo: https://github.com/cnclabs/ICE
  - Demo: https://cnclabs.github.io/ICE/
- IMDB dataset:
  - MovieLens 10/2016 Full dataset.
  - 36,586 movies with plot descriptions and genres.
- Special thanks to Chen Chih-Ming for his help to the development of the API.

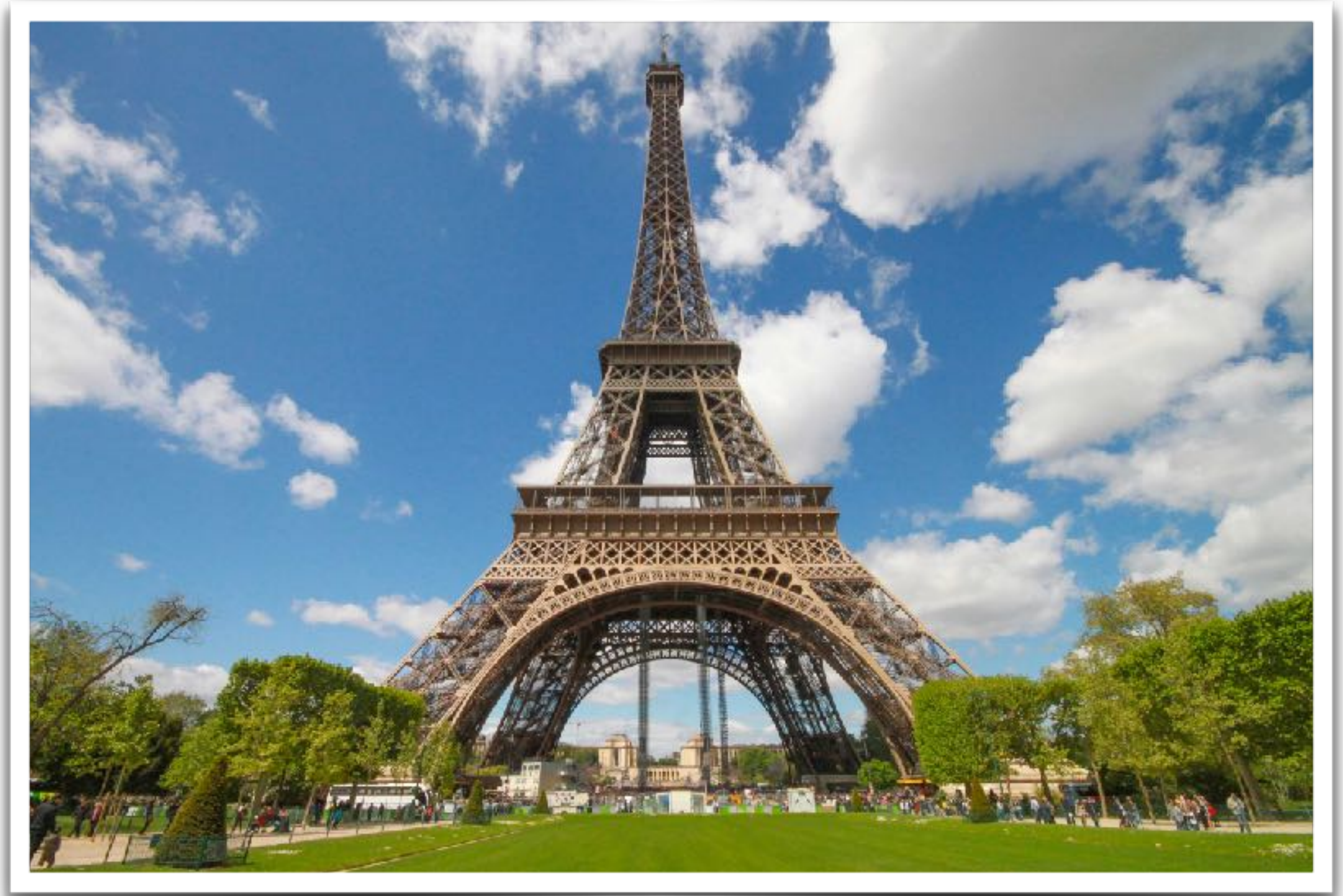
UGSD: User Generated Sentiment Dictionaries from Online Customer Reviews

The 33rd AAAI Conference on Artificial Intelligence
(AAAI'19), Honolulu, 2019.

(full paper, acceptance rate: 16.2%)

https://www.aaai.org/ojs/index.php/AAAI/article/view/3800

Eiffel Tower



User-generated Reviews

Eiffel Tower

Eiffel Tower is an amazing place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well worth paying the extra to get to the top for …

The Eiffel Tower is an overrated land mark and was overpopulated with tourists …

Very disappointing. Lines were crazy, people trying to get you to buy …

User-generated Reviews

Eiffel Tower

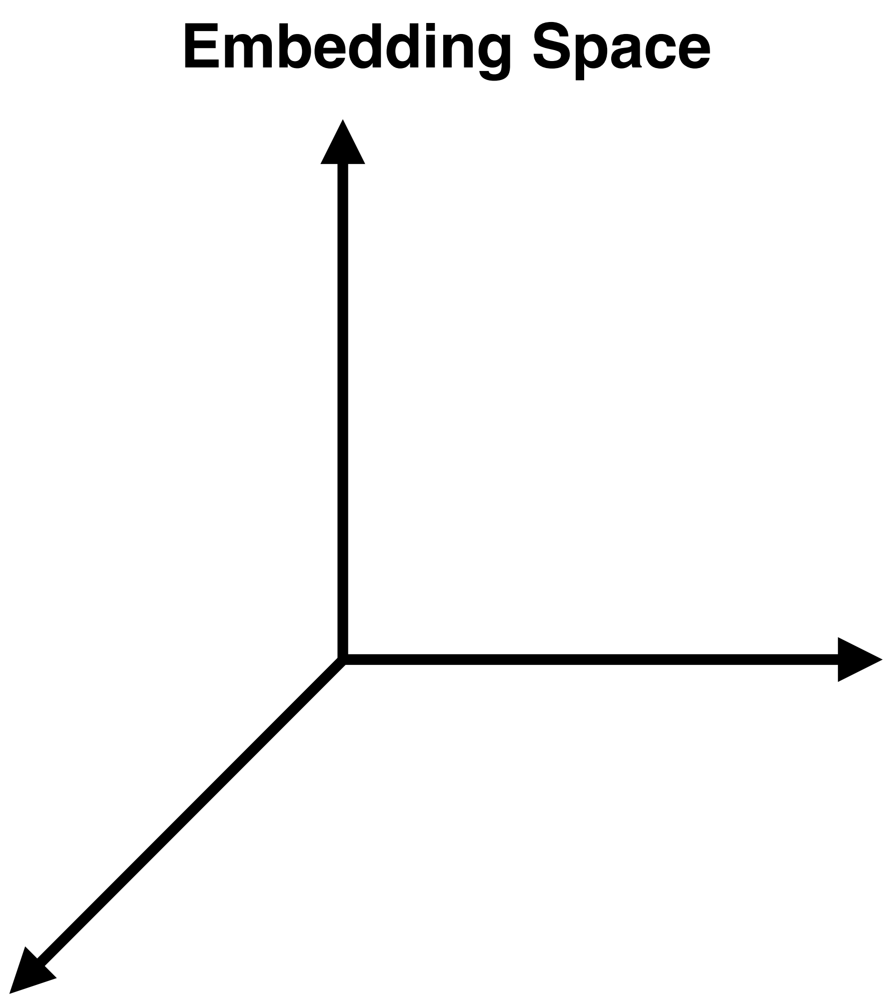
Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Embedding Space



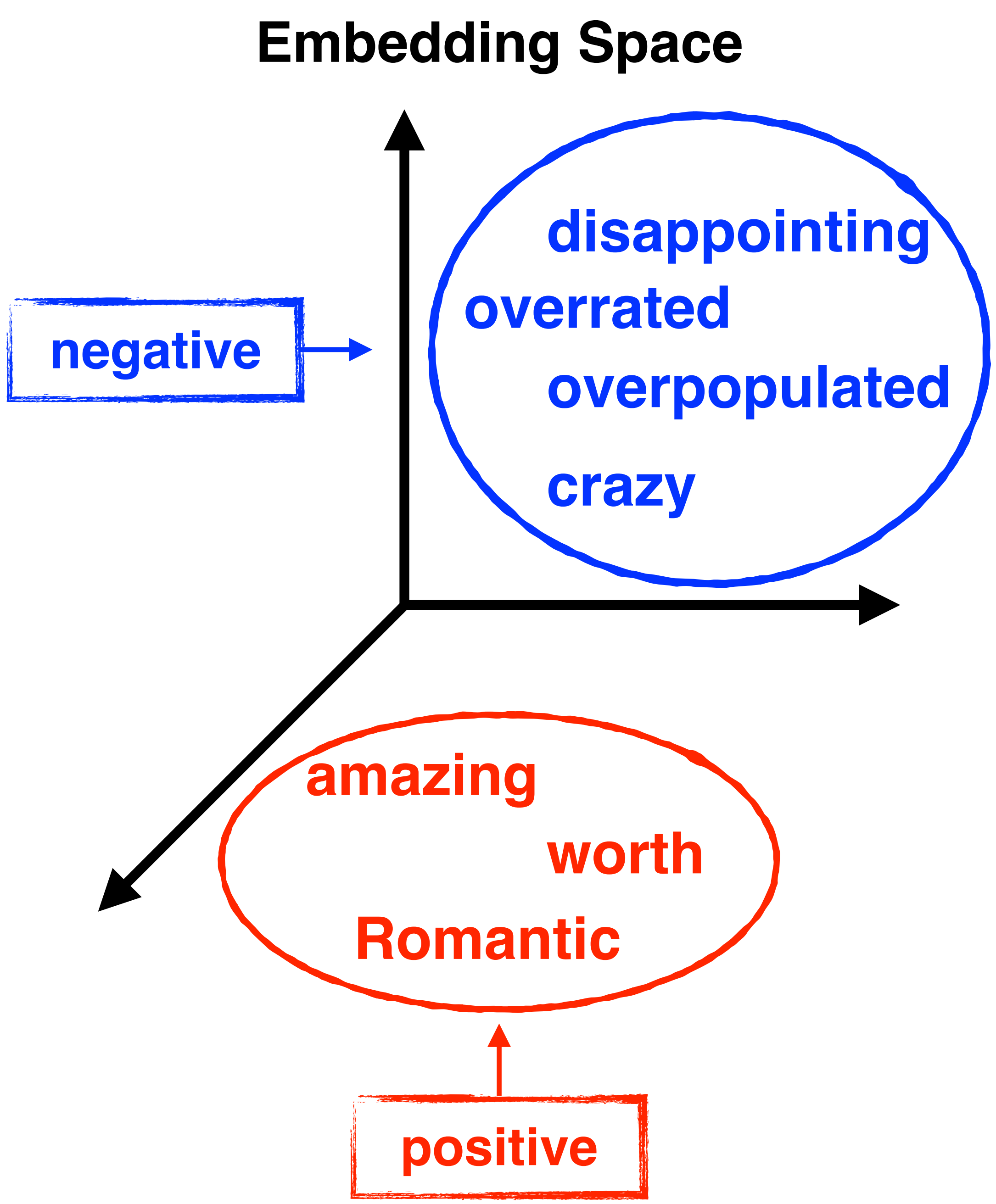
Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Embedding Space



Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Importance of Sentiment Lexicons

- Sentiment analysis and opinion mining
- Sentiment words are domain-specific

TripAdvisor

The hotels in this city are usually too **small** for the whole family to stay overnight.

Amazon

The cellphone is **small** and therefore convenient for people to use it with a single hand.

Our Framework: UGSD

- Construct sentiment lexicons from user-generated reviews
- Features:
  1. Data-driven: require no seed words or external lexicons
  2. Domain-specific: construct domain-specific sentiment lexicons with reviews from different domains
  3. Application scalability: produce representations of the learned sentiment words

Problem Definition

Eiffel Tower

Eiffel Tower is an amazing place to …

Romantic Eiffel Tower. Well worth …

The Eiffel Tower is an overrated land …

Very disappointing. Lines were crazy …

A set of reviews of a certain domain $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$

A rating $r \in \mathcal{R}$ corresponds to each of reviews

A set of entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$

Generate a set of words $\mathcal{L}_r$ corresponding to the rating $r \in \mathcal{R}$

Candidate Word Selection

- Extract adjectives and adverbs as candidates $\mathcal{S} = \{s_1, s_2, \ldots, s_G\}$
- Combine consecutive adverbs and adjectives

Eiffel Tower

Eiffel Towel is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Towel. **Well_worth** paying the extra to get to the top for …

The Eiffel Towel is an **overrated** land mark and was **overpopulated** with tourists …

Very_disappointing. Lines were **crazy**, people trying to get you to buy …

Entity Substitution

- Replace entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ with the rating $r \in \mathcal{R}$

Eiffel Tower

Eiffel Tower is an amazing place to spend at Paris. A must see through out the day …

Romantic **Eiffel Tower**. Well_worth paying the extra to get to the top for …

The **Eiffel Tower** is an overrated land mark and was overpopulated with tourists …

Very_disappointing. Lines were crazy, people trying to get you to buy …

Entity Substitution

- Replace entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ with the rating $r \in \mathcal{R}$

Eiffel Tower

is an amazing place to spend at Paris. A must see through out the day …

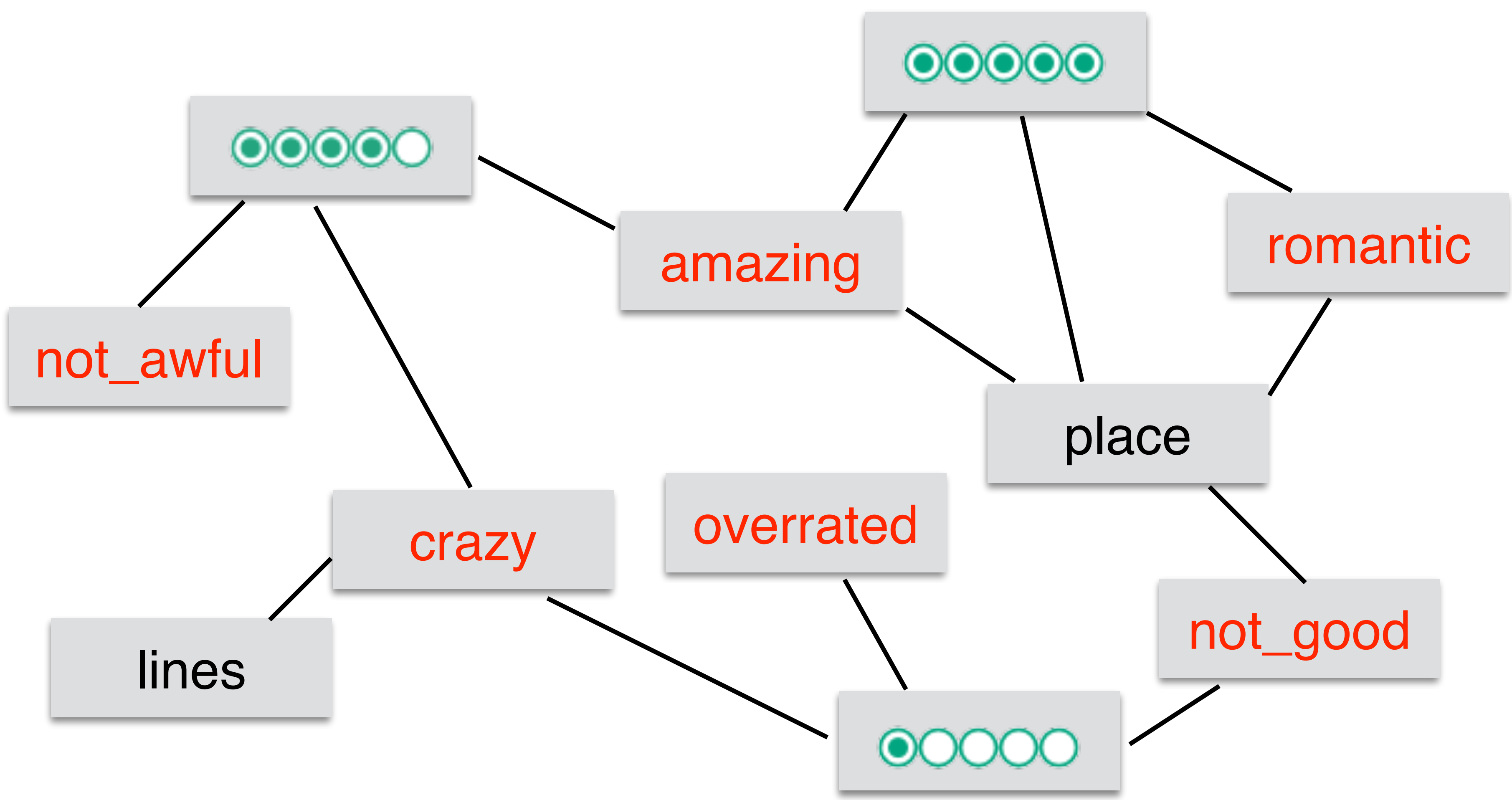
Romantic . Well_worth paying the extra to get to the top for …

The is an overrated land mark and was overpopulated with tourists …

Very_disappointing. Lines were crazy, people trying to get you to buy …

Co-occurrence Proximity Learning

- Construct a *k* co-occurrence network with a predefined window size *k*



Dictionary Construction

- Select the top *N* sentiment words by measuring the cosine similarity

lexicon

1. perfect (0.84)
2. great (0.82)
3. good (0.81)
4. …

…

lexicon

1. annoyed (0.79)
2. unfair (0.74)
3. useless (0.70)
4. …

lexicon

1. not_worth (0.76)
2. poor (0.74)
3. not_great (0.71)
4. …

Real-World Datasets

- Yelp:
  - Round 9 of Yelp dataset challenge
- TripAdvisor dataset:
  - Top 25 cities in 2016 and top 20 attractions or tours of each city
- Amazon dataset: (Wang, et al., 2010)
  - 6 categories of electronic supplies and top 20 products of each category

Comparison with Yelp Dictionary

- Compare Yelp dictionaries with the state-of-the-art Yelp dictionaries (Reschke, et al., 2013)

		Positive					Negative			
		# word	P	R	F1		# word	P	R	F1
	NLTK	2,006	0.196	0.275	0.229		4,783	0.072	**0.607**	0.129
	MPQA	2,304	0.198	0.318	0.244		4,152	0.079	0.579	0.139
	SentiWordNet	14,712	0.039	0.395	0.071		10,751	0.015	0.288	0.029
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5}$	594	**0.352**	0.146	0.206	$\mathcal{L}_{r_1}$	1,112	0.161	0.314	0.213
	$\mathcal{L}_{r_{45}}$	1,125	0.332	0.260	0.292	$\mathcal{L}_{r_{12}}$	1,901	0.140	0.467	0.215
	$\mathcal{L}_{r_{345}}$	1,685	0.315	0.369	0.340	$\mathcal{L}_{r_{123}}$	2,461	0.119	0.512	0.193
$\mathcal{G}_{z>0.6}(\cdot)$	$\mathcal{L}_{r_5}$	1,309	0.349	0.318	0.333	$\mathcal{L}_{r_1}$	534	**0.281**	0.263	0.272
	$\mathcal{L}_{r_{45}}$	1,860	0.322	0.417	**0.363**	$\mathcal{L}_{r_{12}}$	773	0.247	0.335	**0.284**
	$\mathcal{L}_{r_{345}}$	2,113	0.296	**0.436**	0.353	$\mathcal{L}_{r_{123}}$	990	0.202	0.351	0.256

Sentiment Classification

- Conduct binary sentiment classification on reviews for three datasets

		Yelp			TripAdvisor			Amazon		
		# word	F1	Acc	# word	F1	Acc	# word	F1	Acc
	NLTK	6,787	0.762	0.697	6,787	**0.759**	0.699	6,787	0.766	0.707
	MPQA	6,450	0.708	0.601	6,450	0.701	0.608	6,450	0.716	0.616
	SentiWordNet	24,123	0.675	0.534	24,123	0.670	0.520	24,123	0.685	0.551
	Stanford Yelp	2,005	0.682	0.534	2,005	0.686	0.544	2,005	0.679	0.530
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5} \cup \mathcal{L}_{r_1}$	1,524	0.733	0.755	1,888	0.664	0.679	717	0.744	0.727
	$\mathcal{L}_{r_{45}} \cup \mathcal{L}_{r_{12}}$	2,692	0.771	**0.777**	3,428	0.746	**0.753**	1,566	0.763	0.755
$\mathcal{G}_{z>1.2}(\cdot)$	$\mathcal{L}_{r_5} \cup \mathcal{L}_{r_1}$	364	0.784	0.758	710	0.726	0.630	189	**0.801**	**0.782**
	$\mathcal{L}_{r_{45}} \cup \mathcal{L}_{r_{12}}$	451	**0.792**	0.762	1,060	0.736	0.650	346	0.800	0.772

Entity Ranking

- Entity ranking performance

		TripAdvisor			Amazon		
		# word	NDCG@5	NDCG@10	# word	NGCG@5	NDCG@10
Frequency		-	0.610	0.664	-	0.494	0.623
NLTK		1,071	0.556	0.632	595	0.603	0.659
MPQA		1,294	0.562	0.641	710	0.571	0.654
SentiWordNet		4,522	0.442	0.530	2,207	0.543	0.574
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5}$	207	**0.794**	**0.818**	258	**0.635**	**0.712**
	$\mathcal{L}_{r_{45}}$	745	0.669	0.724	493	0.549	0.641
	$\mathcal{L}_{r_{345}}$	1,626	0.654	0.698	995	0.574	0.655
$\mathcal{G}_{z>1.2}(\cdot)$	$\mathcal{L}_{r_5}$	288	0.782	0.807	51	0.606	0.695
	$\mathcal{L}_{r_{45}}$	569	0.735	0.770	114	0.515	0.631
	$\mathcal{L}_{r_{345}}$	895	0.719	0.751	221	0.515	0.627

Amazon Lexicons

Top	$\mathcal{L}_{r_5}$	θ_s^{r5}	$\mathcal{L}_{r_4}$	θ_s^{r4}
1	wonderful wonderfully	0.599	not_perfect	0.695
2	fantastic fantastically	0.538	overall	0.600
3	awesome	0.536	standalone	0.525
4	amazing amazingly	0.532	nice nicely	0.503
5	really_great	0.526	good	0.469
6	great greatly	0.503	almost_perfect	0.449
7	lovely loving	0.428	lightest	0.312
8	excellent excellently excelent excellant	0.406	far_satisfied	0.290
9	best	0.369	little	0.284
10	absolutely_wonderful	0.347	starter	0.281
11	exellent	0.319	great greatly	0.265
12	happy	0.315	pretty_happy	0.257
13	really_loving	0.297	solid solidly	0.256
14	smart	0.290	graphically_intense	0.238
15	ever	0.271	not_primary	0.220
16	absolute absolutely absolutly	0.263	uncertain	0.219
17	totally_satisfied	0.258	not_expensive	0.199
18	bought	0.251	still_amazing	0.194
19	beatiful	0.242	darn darned	0.165
20	perfect perfectly	0.225	not_smart	0.163

Amazon Lexicons

Disappointed. The phone is **not new**, it is a used phone.

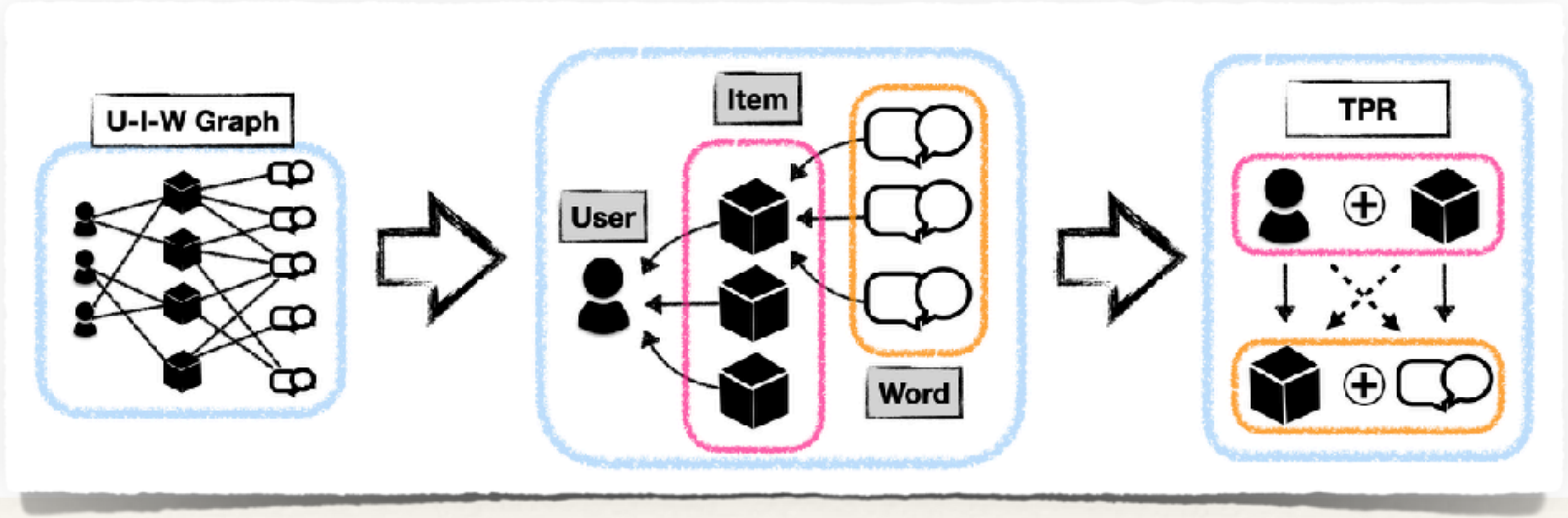
$\mathcal{L}_{r_3}$	θ_s^{r3}	$\mathcal{L}_{r_2}$	θ_s^{r2}	$\mathcal{L}_{r_1}$	θ_s^{r1}
okay	0.813	unfortunate unfortunately	0.785	extremely_disappointed	0.769
ok	0.605	not_good	0.626	worthless	0.740
alright	0.583	disappointed disappointing	0.579	not_new	0.631
not_bad	0.521	not_waterproof	0.542	worse	0.609
dumb	0.517	really_disappointed really_disappointing	0.516	far_worst	0.594
not_great	0.418	unreliable	0.508	unacceptable	0.589
decent decently	0.399	dissapointed dissapointing	0.508	totally_useless	0.583
temporary	0.386	not_smart	0.480	useless	0.578
otherwise	0.375	overrated	0.458	faulty	0.576
pretty_decent	0.346	sad sadly	0.409	not_acceptable	0.568
not_smooth	0.297	not_happy not_happier	0.400	lemon	0.531
bland	0.290	unbearable	0.389	dissatisfied	0.527
not_happy not_happier	0.283	not_worst	0.386	not_happy not_happier	0.524
not_crazy	0.276	absolutely_terrible	0.370	apparent apparently	0.514
really_annoying	0.271	unhappy	0.367	defective	0.512
beloved	0.265	astonishing	0.359	miserable miserably	0.509
fully_capable	0.264	ongoing	0.351	unusable unused	0.488
really_excellent	0.248	still_slow	0.349	unhappy	0.487
wise	0.247	not_worth	0.342	ashamed	0.483
inaccurate	0.236	frustrated frustrating	0.339	completely_dead	0.472

Short Recap

- Propose a representation learning framework for constructing sentiment dictionaries from user reviews
  - Data-driven
  - Domain-specific
  - Application scalability
- Code & Datasets: github.com/cnclabs/UGSD

Our more recent work

- TPR: Text-aware Preference Ranking for Recommender Systems, CIKM full paper, 2020.
    - https://github.com/cnclabs/codes.tpr.rec



Thanks for Your Listening!