FIRE 🔥 Semantic Field of Words Represented as Non-Linear Functions

University of Tokyo

Xin Du & Kumiko Tanaka-Ishii

Mathematical Representation of Words, Sentences

Embeddings in a space

Important basis of machine learning for natural language processing

Typically a linear vector space
Word2Vec, BERT

- Similar words must be mapped to similar embeddings
- Compositionality
- Polysemy

Mathematical Representation of Words, Sentences

Embeddings in a space

Important basis of machine learning for natural language processing

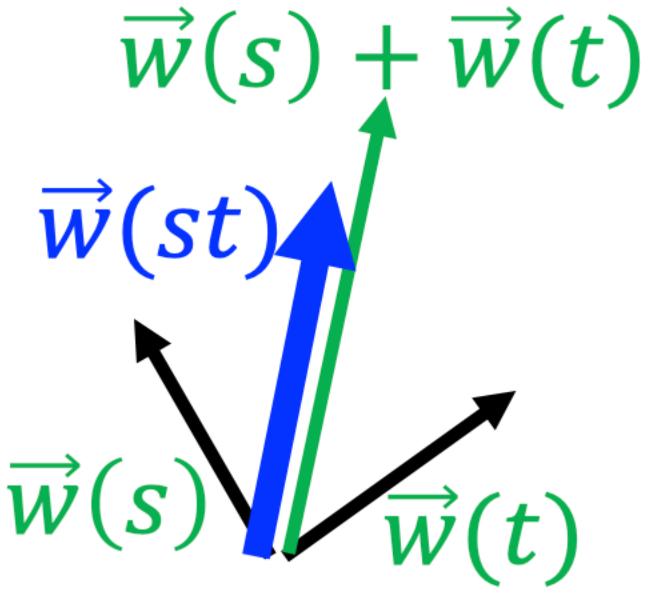
Typically a linear vector space
Word2Vec, BERT

- Similar words must be mapped to similar embeddings
- Compositionality linear quality

$$\vec{w}(\text{coffee}) + \vec{w}(\text{cup}) \approx \vec{w}(\text{coffee cup})$$

- Polysemy non-linear quality

bank: financial bank vs. river bank



Adding polysemy often destroys compositionality

Research question:

How to incorporate linearity and non-linearity quality?

Comparison with Previous Work

D: dimension of representation
K: max number of polysemy
L: number of neural layers

method	Non-Contextual	Compositionality	Polysemy	Interpretability	N (# of parameters)	Complexity $\mathrm{sim}(w_1, w_2)$
			Vectoral representation			
Word2Vec (2013)	√	√	×	×	D	$\mathcal{O}(N)$
GloVe (2014)	√	√	×	×	D	$\mathcal{O}(N)$
BERT-large (2019)	×	√	√	×	$D = 1024$	high
			Random-variable representations			
Word2Gauss/S (2014)	√	×	×	×	$D + 1$	$\mathcal{O}(N)$
Word2Gauss/D (2014)	√	×	×	×	$2D$	$\mathcal{O}(N)$
Word2GM/S (2017)	√	×	√	×	$(D + 2)K$	$\mathcal{O}(KN)$
Word2GM/D (2017)	√	×	√	×	$(2D + 1)K$	$\mathcal{O}(KN)$
Word2Cloud (2019)	√	×	√	√	$K = 64$	$\mathcal{O}(N^2)$
CMD (2020)	√	nonlinear	√	×	$K = 200, 400$	$O(N^2)$
			Our semantic-field representations			
FIRE (2022)	√	√	√	√	$(2D+1)L+(D+1)K$	$\mathcal{O}(KL)$
FIRE/m (2022)	√	√	√	√	$(2D + 1)L + DK$	$\mathcal{O}(KL)$

FIRE : Representation of Words in a Functional Space

$$w_i = [\mu_i, f_i(s)]$$

A space $s \in S$

location function

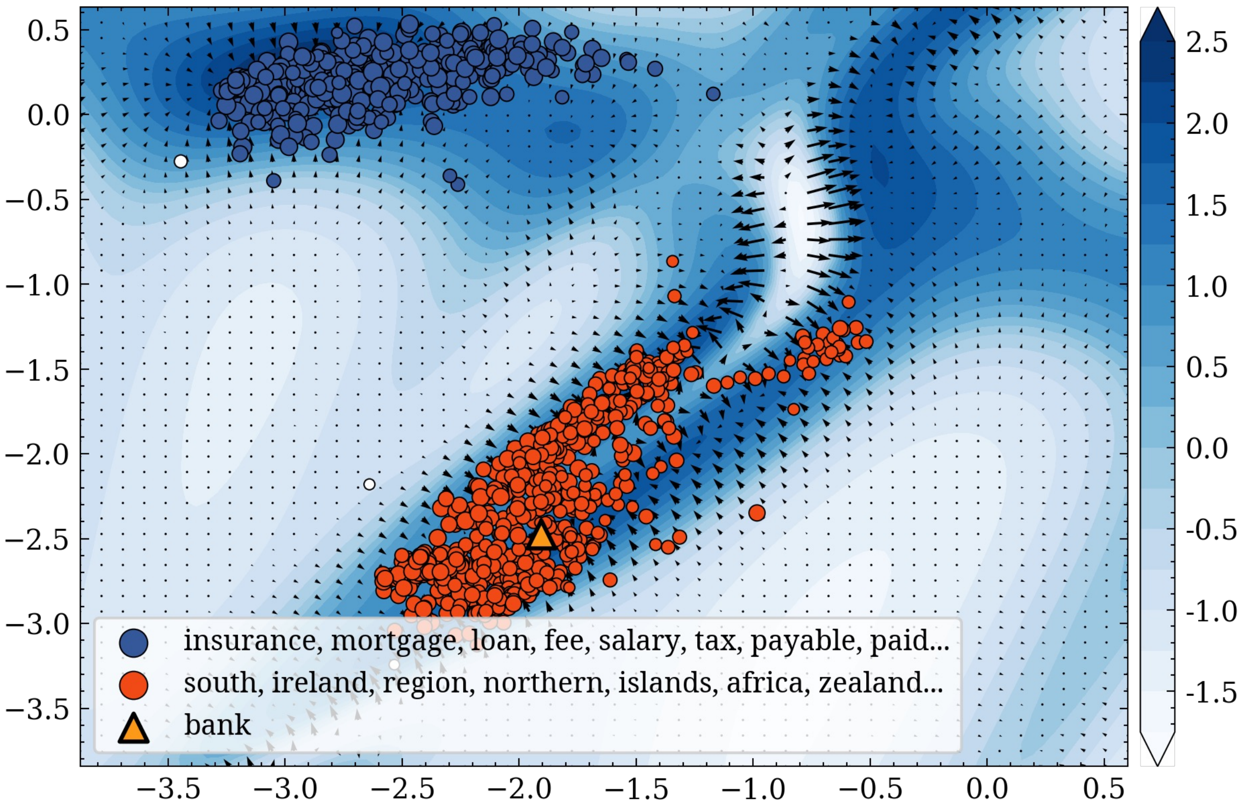
Representing word context

measure neural network $S \to \mathbb{R}$

$$\mu_i \equiv \delta(s_i)$$

s_i is the location of word w_i

"bank" $f_{bank}(s)$ and s_{bank}



FIRE : Representation of Words in a Functional Space

$$w_i = [\mu_i, f_i(s)]$$

A space $s \in S$

location — Representing word

function — context

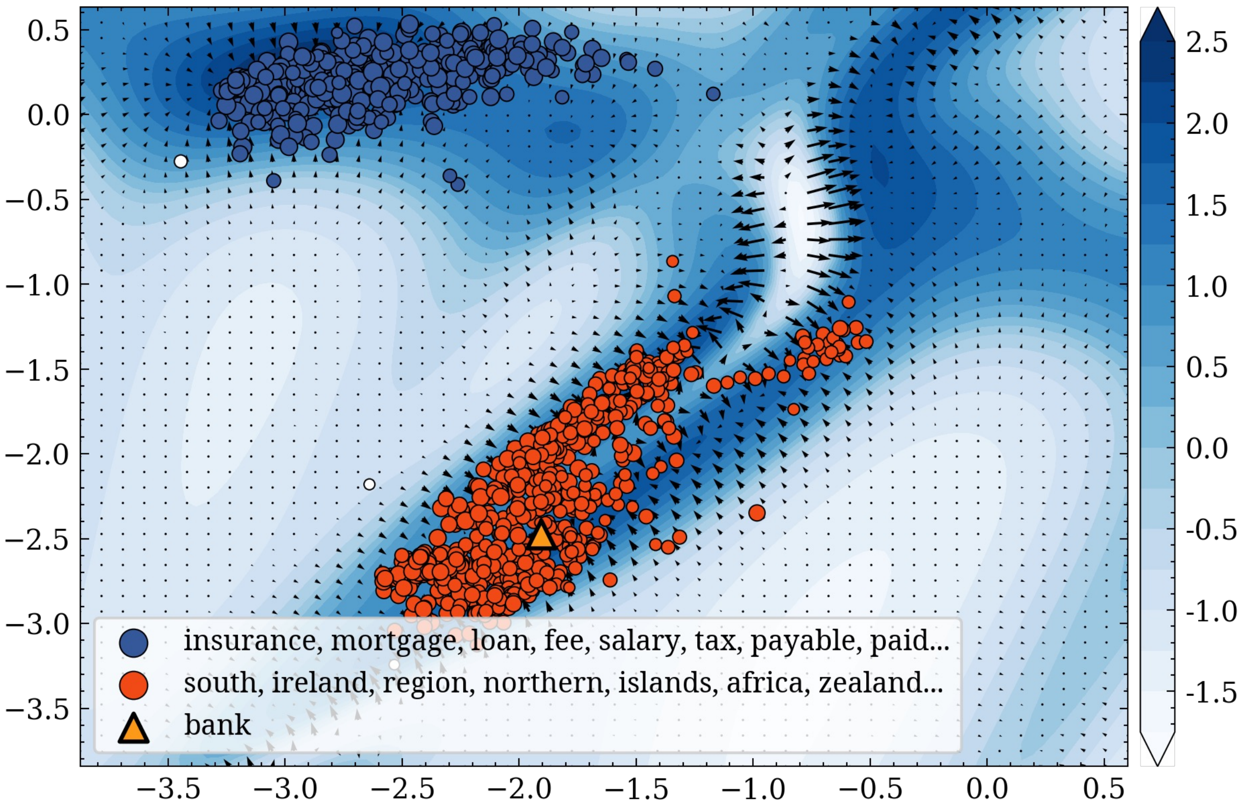
measure — neural network $S \to \mathbb{R}$

$\mu_i \equiv \delta(s_i)$ — s_i is the location of word w_i

The interaction of w_j in the context of w_i

$$\int f_j \mathrm{d}\mu_i = f_j(s_i)$$

“bank” $f_{bank}(s)$ and s_{bank}



Similarity function $\mathrm{sim}(w_i, w_j) \equiv \int f_i \,\mathrm{d}\mu_j + \int f_j \mathrm{d}\mu_i$

Compositionality: Addition of $f(s)$

Polysemy: Shape of $f(s)$ + number of locations K

Simple and Natural Extentions

- Polysemy by K locations per word

$$\mu \equiv \sum_{k=1}^{K} m^{(k)} \delta(s^{(k)})$$

$m^{(k)}$: weights (acquired by training)

- Sentence Representation

$$\Gamma = [w_1, \ldots, w_n] \qquad [\mu_1, f_1(s)], \ldots, [\mu_n, f_n(s)]$$

$$\mu = \sum_{i=1}^{n} \gamma_i \mu_i, \quad f(s) = \sum_{i=1}^{n} \gamma_i f_i(s)$$

γ_i : weights (we use SIF)

$$\gamma = [\gamma_1, \ldots, \gamma_n]^{\mathrm{T}} \qquad \gamma' = [\gamma'_1, \ldots, \gamma'_{n'}]^{\mathrm{T}}$$

$$\mathrm{sim}(\Gamma, \Gamma') = \int \Big(\sum_{i=1}^{n} \gamma_i f_i\Big)\, \mathrm{d}\Big(\sum_{j=1}^{n'} \gamma'_j \mu'_j\Big) + \int \Big(\sum_{j=1}^{n'} \gamma'_j f'_j\Big)\, \mathrm{d}\Big(\sum_{i=1}^{n} \gamma_i \mu_i\Big) = \gamma^{\mathrm{T}} \Sigma \gamma'$$

Implementation of FIRE via Skipgram

$$\min \sum_{w_i, w_p, w_n} \sigma\big(-\mathrm{sim}(w_i, w_p)\big) + \sigma\big(\mathrm{sim}(w_i, w_n)\big)$$

σ : soft-plus function.

w_p positive samples : Words that co-occur with w_i

w_n negative samples : Words that do not co-occur with w_i

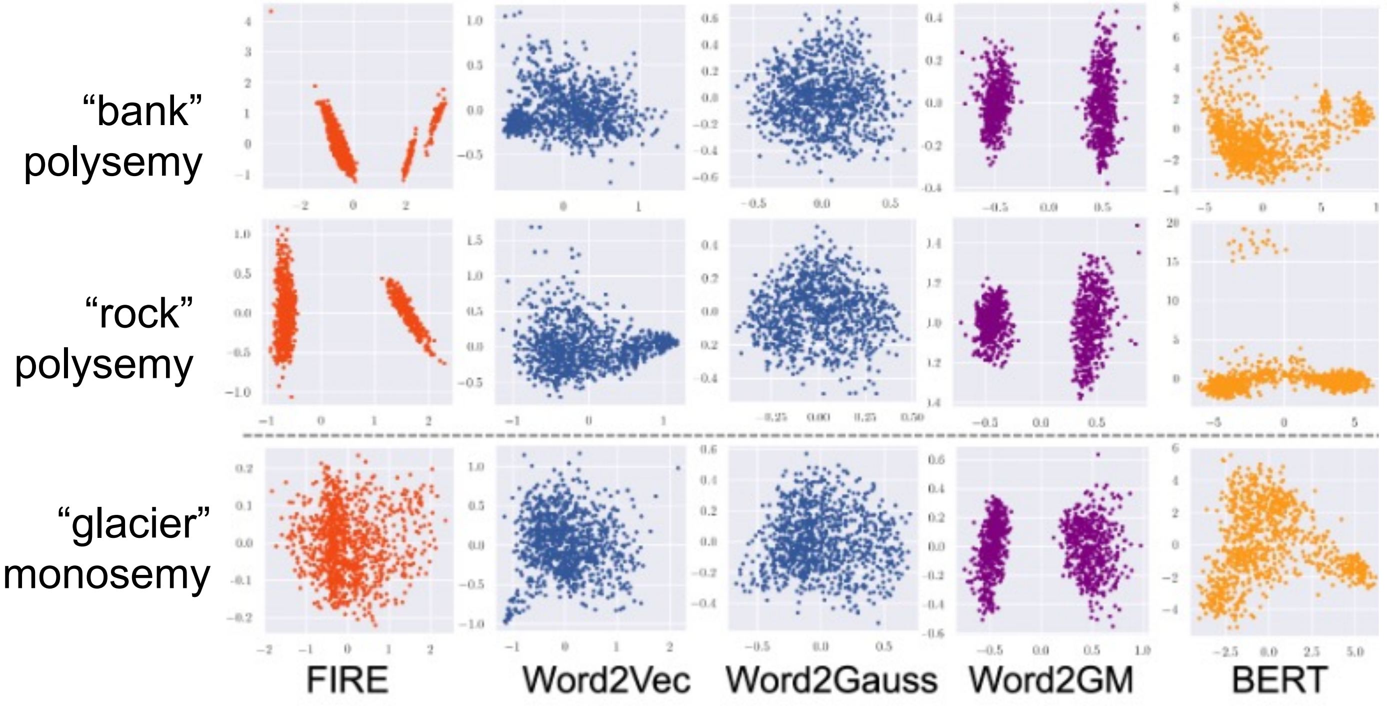
$w_i = [\mu_i, f_i(s)]$ train f_i (s_i) and s_i for every w_i

location function

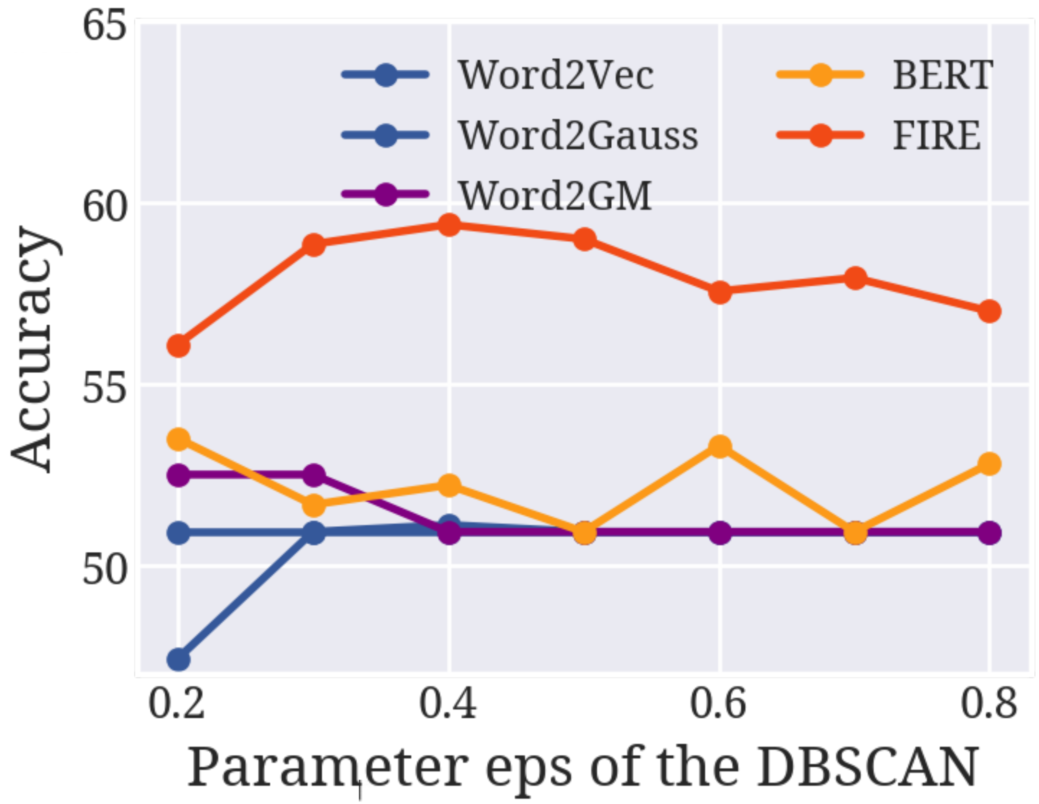
We used MLPlanar

Evaluation of FIRE

1. Word similarity benchmarks: FIRE competes well with SOTA
2. Sentence similarity benchmarks: less than BERT, compete with Word2Vec
3. Polysemy / monosemy classification



2d-PCA classification of representation
Polysemy : two clouds
Monosemy: one cloud
Only FIRE could achieve this distinction



542 words ← Wordnet dictionary
Classification: “polysemy” vs. “monosemy”
Only FIRE achieved better than a chance level

Thank you