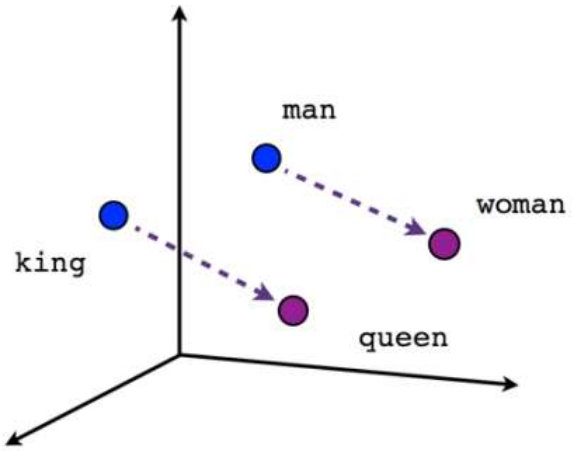
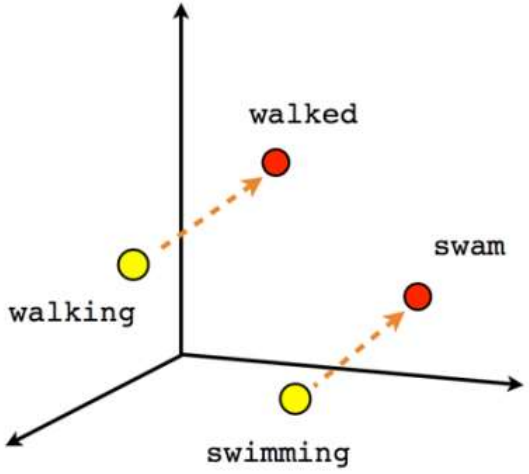
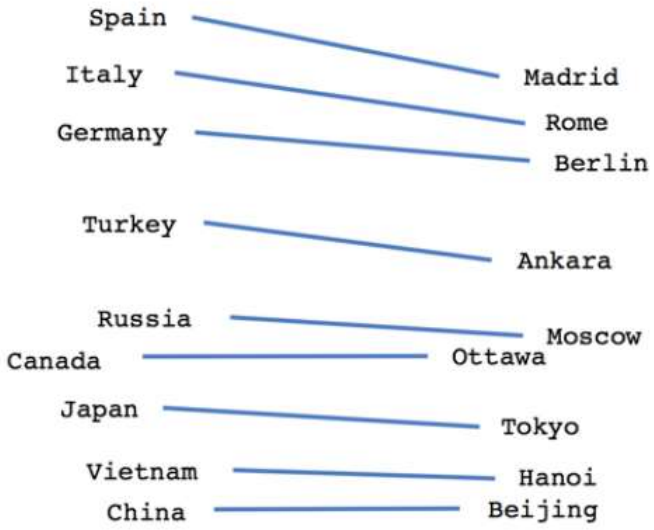
- Word embeddings



Male-Female

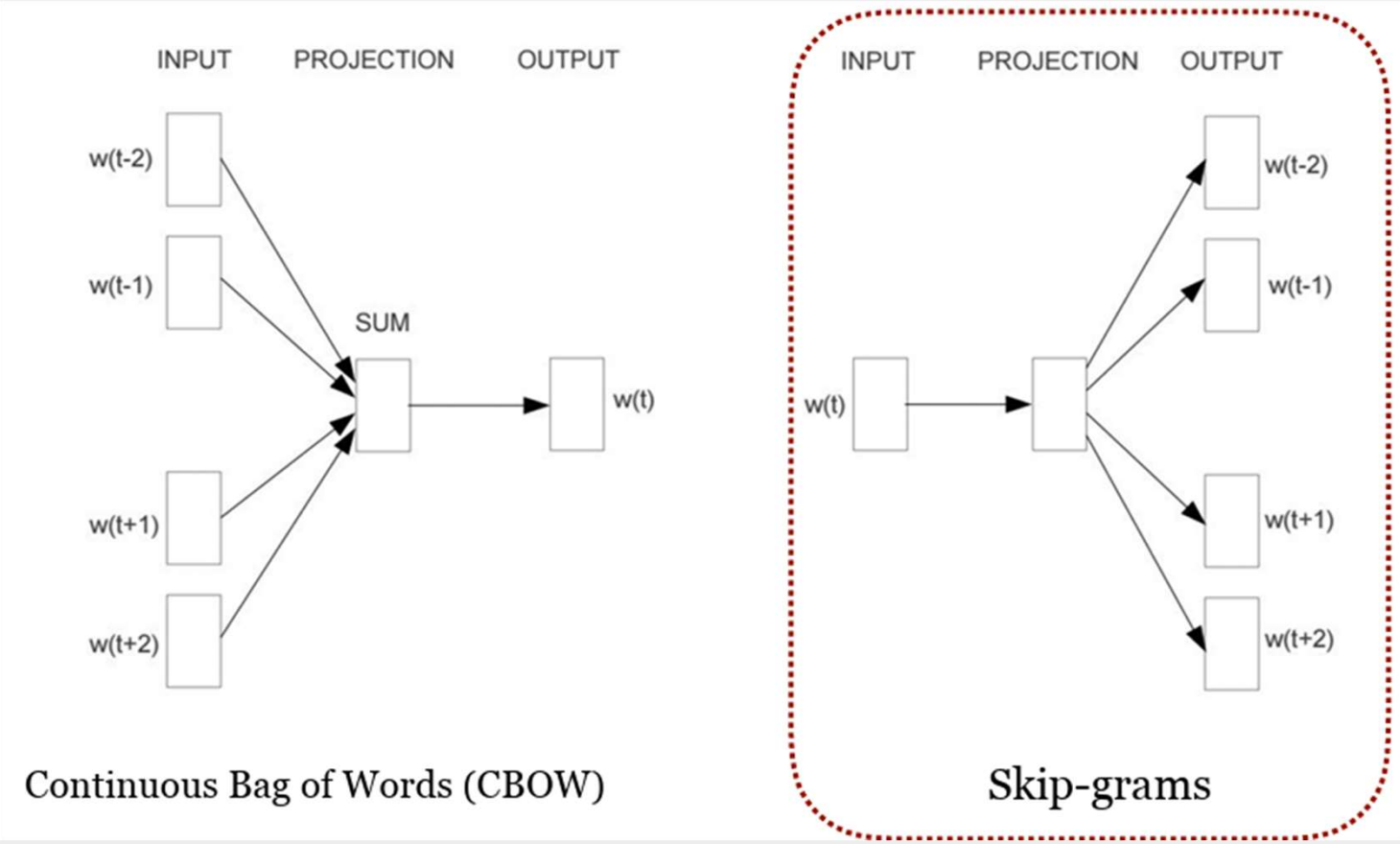


Verb tense

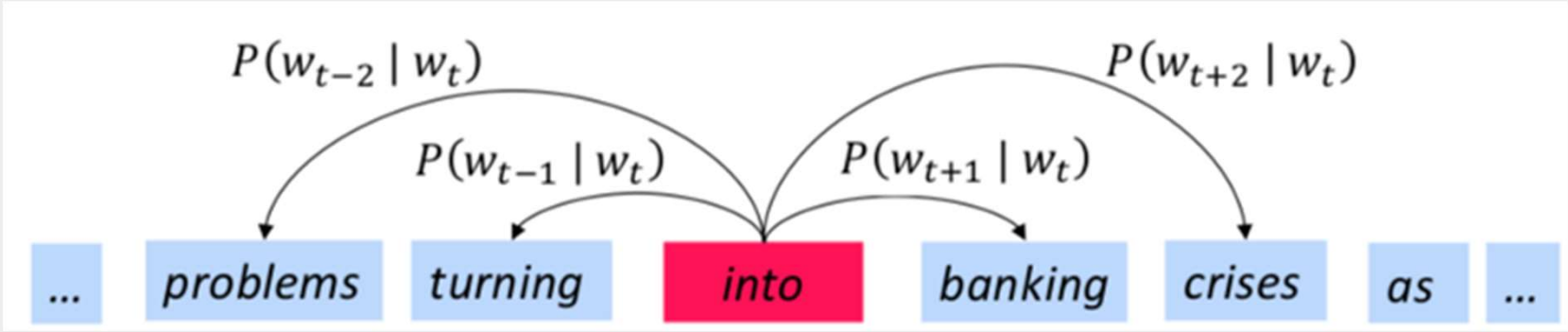


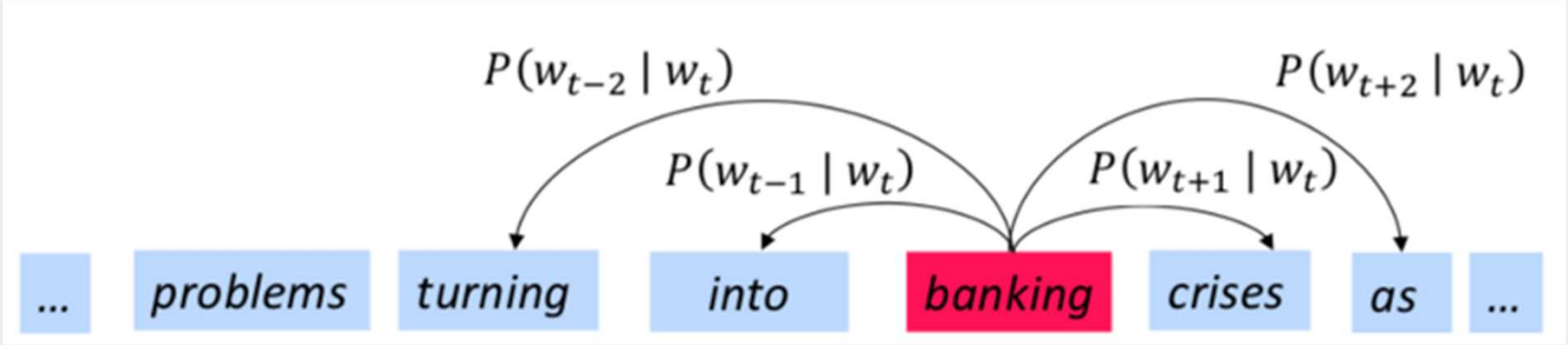
Country-Capital

- Word embeddings – Word2Vec



- Skip-gram


$P(w_{t-2} \mid w_t)$
$P(w_{t-1} \mid w_t)$
$P(w_{t+1} \mid w_t)$
$P(w_{t+2} \mid w_t)$
...
problems
turning
into
banking
crises
as
...


$P(w_{t-2} \mid w_t)$
$P(w_{t-1} \mid w_t)$
$P(w_{t+1} \mid w_t)$
$P(w_{t+2} \mid w_t)$
...
problems
turning
into
banking
crises
as
...

- Función objetivo
  - Para cada posición t = 1,2…T predecir las palabras del contexto (tamaño m) dada la palabra w_j

Funcion de verosimilitud

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m, j \neq 0} P(w_{t+j} \mid w_t; \theta)$$

$$J(\theta) = -\frac{1}{T} \log \mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \leq j \leq m, j \neq 0} \log P(w_{t+j} \mid w_t; \theta)$$

- Cómo calcular esa probabilidad? $P(w_{t+j} \mid w_t; \theta)$

$\mathbf{u}_i \in \mathbb{R}^d$ El embedding para la palabra i

$\mathbf{v}_{i'} \in \mathbb{R}^d$ El embedding del contexto de i

$$P(w_{t+j} \mid w_t) = \frac{\exp(\mathbf{u}_{w_t} \cdot \mathbf{v}_{w_{t+j}})}{\sum_{k \in V} \exp(\mathbf{u}_{w_t} \cdot \mathbf{v}_k)}$$

- Luego la cantidad de parámetros es

$$\theta = \{\{\mathbf{u}_k\}, \{\mathbf{v}_k\}\}$$

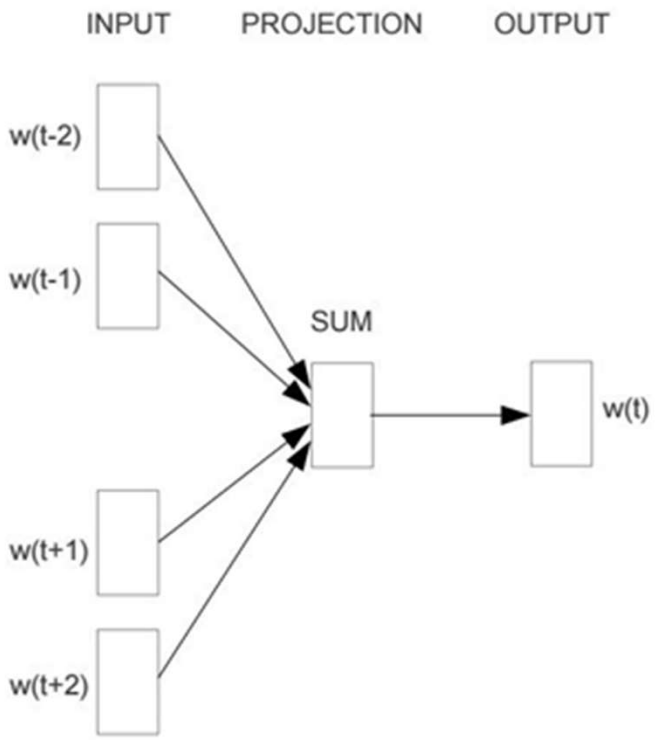
- Y se debe optimizar

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0} \log P(w_{t+j} \mid w_t; \theta) \qquad \nabla_\theta J(\theta) = ?$$

- Usando descenso por gradiente estocástico

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta J(\theta)$$

- Continuous Bag of Words



$$L(\theta) = \prod_{t=1}^{T} P\left(w_t \mid \{w_{t+j}\}, -m \leq j \leq m, j \neq 0\right)$$

$$\bar{\mathbf{v}}_t = \frac{1}{2m} \sum_{-m \leq j \leq m, j \neq 0} \mathbf{v}_{t+j}$$

$$P(w_t \mid \{w_{t+j}\}) = \frac{\exp(\mathbf{u}_{w_t} \cdot \bar{\mathbf{v}}_t)}{\sum_{k \in V} \exp(\mathbf{u}_k \cdot \bar{\mathbf{v}}_t)}$$

Criterio	Skip-gram	CBOW
Velocidad	Más lento (una predicción por palabra)	Más rápido (una sola predicción)
Rendimiento con corpus pequeño	Mejor: más preciso con datos escasos	Peor: promedia y pierde matices
Palabras raras	Se aprende mejor	Tienden a ser subrepresentadas
Captura mejor las relaciones	Sí (porque cada contexto se procesa por separado)	No tan bien (suma del contexto)
Simplicidad de implementación	Ligeramente más complejo	Más simple