PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

SoCVis
PUC SOCIAL COMPUTING & VISUALIZATION GROUP

word2vec y algunas aplicaciones

Jorge Schellman Sepúlveda

Profesor Supervisor: Denis Parra

Language Model (LM)

- Modelo para cuantificar la co-ocurrencia de secuencias de palabras en texto
  - “El perro se comió mi tarea” → s = 1000
  - “El perro se comió mi comida” → s = 3

Probabilistic Language Model (PLM)

- Score := Probabilidad
- Ej.: n-grams

Neural Probabilistic Language Model (NPLM)

- LMs que usan enfoques de NN para obtener las probabilidades
- Ej.: Bengio (2003), Mikolov (2013) aka word2vec

n-grams

- Def.: secuencia de **n** consecutivas palabras
- Descomposición en bigramas de “Me gustan los perros”:
  - [“Me gustan”, “gustan los”, “los perros”]
- Meta: $P(w_t, w_{t-1}, w_{t-2}, \ldots, w_{t-(n-1)})$
- Nos limitamos a la historia relevante:
  - $P(w_t \mid w_{t-1:1}) = P(w_t \mid w_{t-1:t-(n-1)})$
- Modelos n-grams son entrenados con conteos de frecuencia en corpus grandes de texto (modelos multinomiales)

n-grams

- ✓ Simple
- ✓ Rendimiento razonable
- × Número de parámetros crece exponencialmente con el tamaño del contexto (**<u>n</u>**-gram)

(Representaciones discretas)

- × No podemos generalizar a instancias no observadas: n-grams poco/nada frecuentes
- × Todos los valores discretos son igualmente parecidos. ¿Vecindad?
  - “Restaurant” y “Coffee-Shop” son sólo “conceptos”, no “conceptos estrechamente relacionados”

Representaciones continuas

- “Distributional Hypothesis” (Harris, 1954): palabras con contextos similares tienen significados similares (*linguistic items with similar distributions have similar meanings*).
- Modelos NPLM usan vectores “continuos” para representar palabras
- Un NPLM puede usar instancias de entrenamiento como:

“Iré a cocinar carne”

para obtener información semántica de:

“Iré a cocinar yuca”

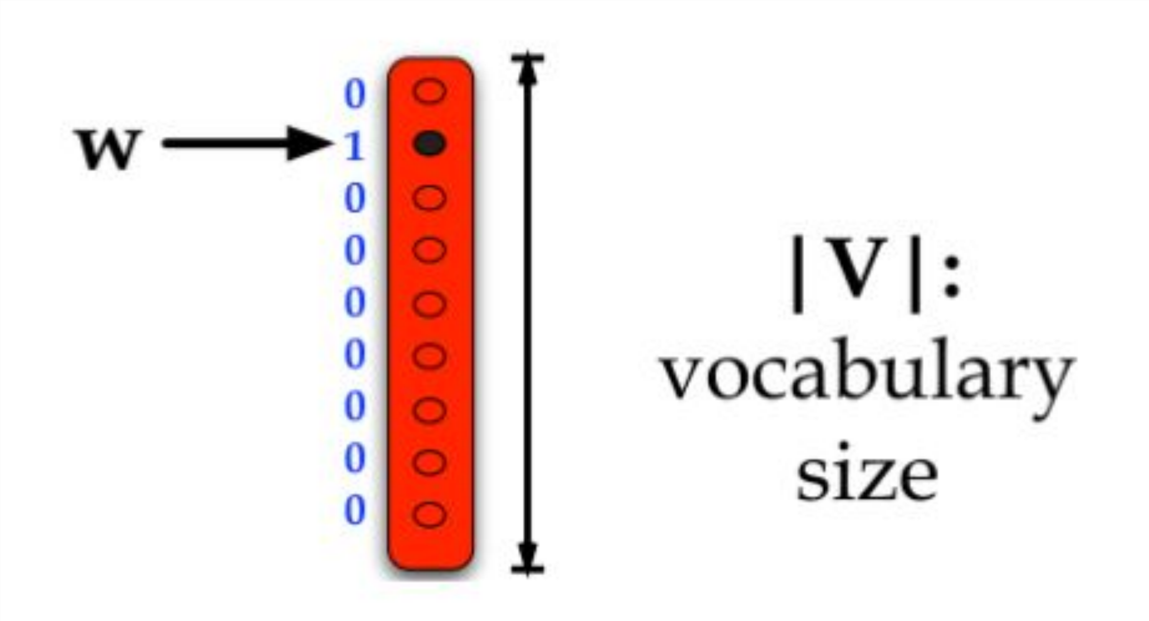
Bengio et al. (NIPS, 2000; JMLR, 2003)

Proponen uno de los primeros NPLM con una feedforward NN

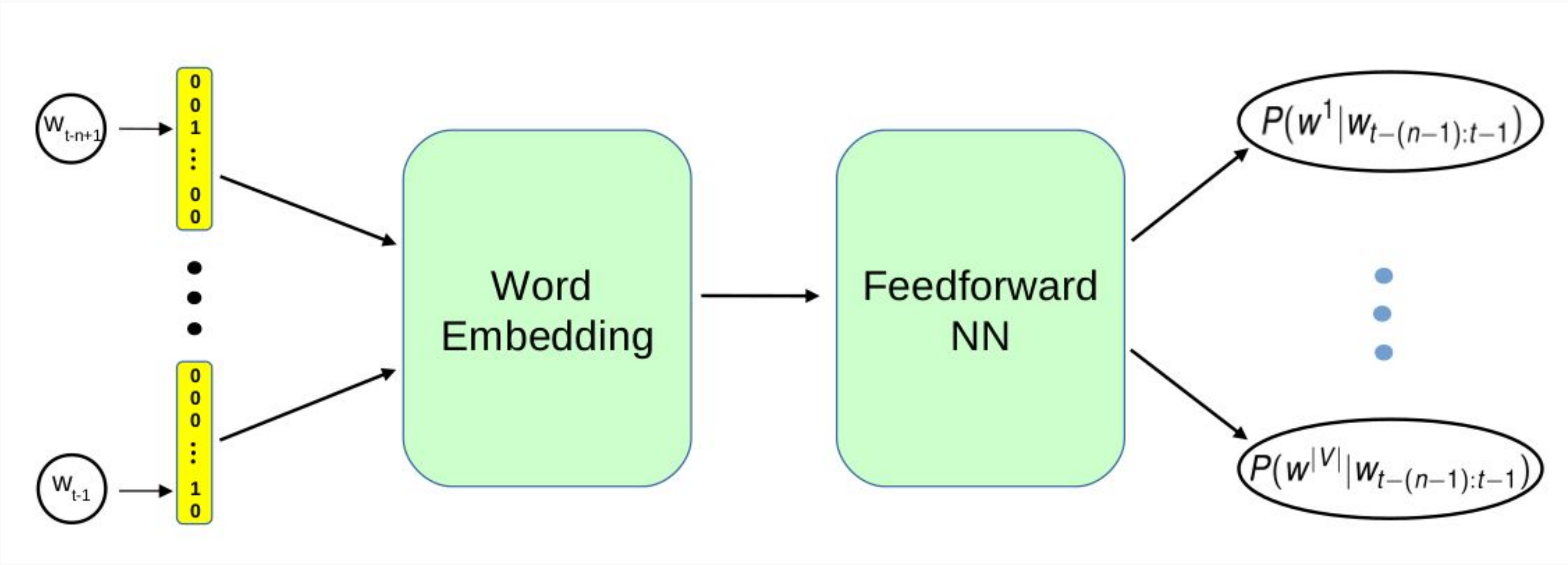
- Meta 1: aprender un modelo para predecir $P(w \mid h(w))$
- Meta 2: aprender un ***word embedding***

Bengio et al. (NIPS, 2000; JMLR, 2003)

- one-hot encoding

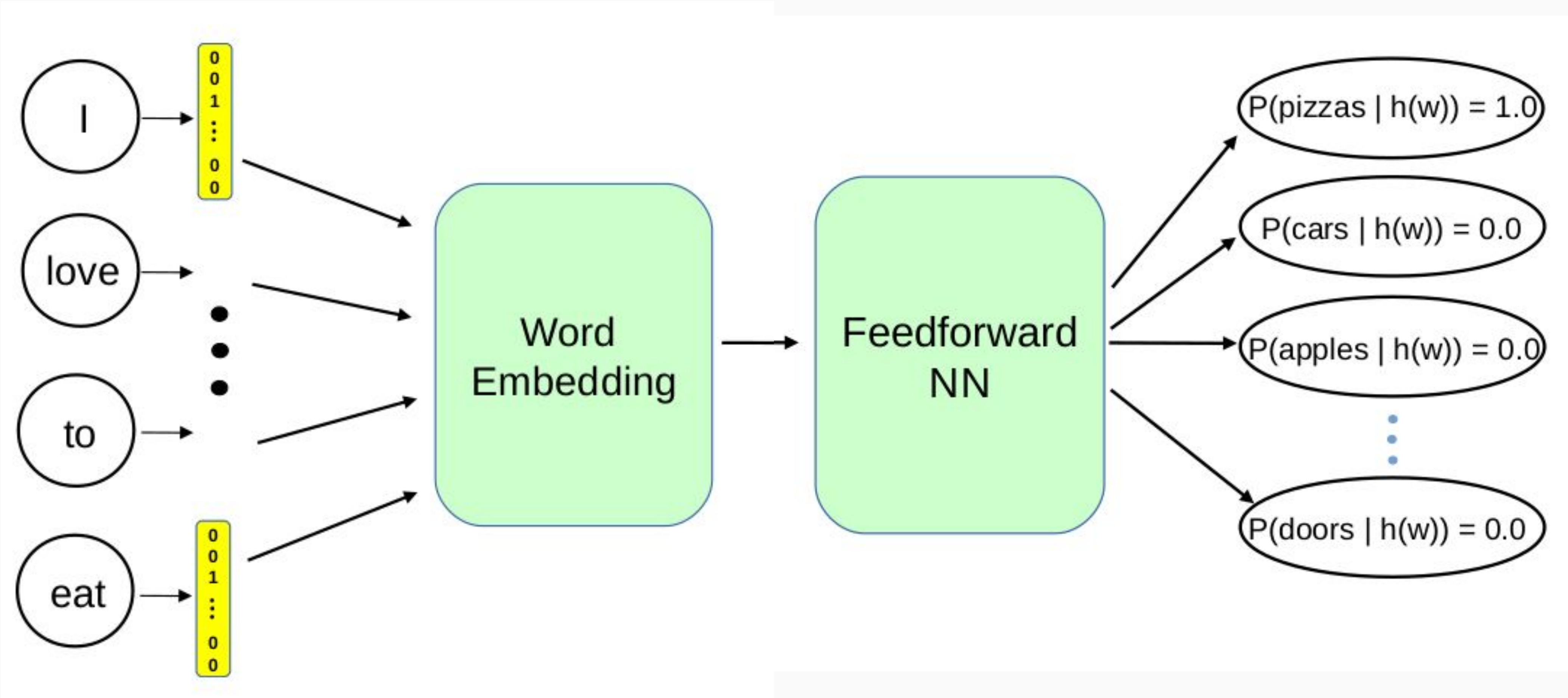


Bengio et al. (NIPS, 2000; JMLR, 2003)



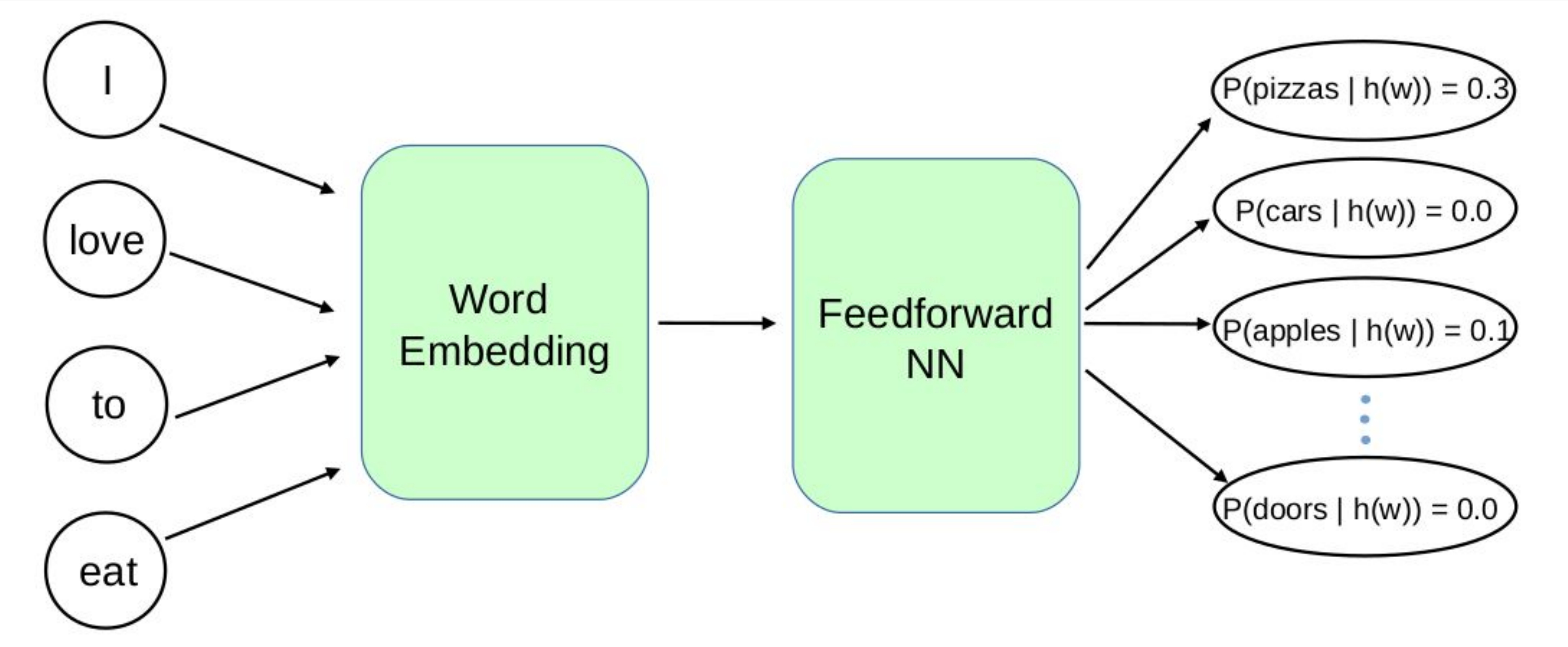
Bengio et al. (NIPS, 2000; JMLR, 2003)

- Instancia de entrenamiento etiquetada (x,y):
  - x: I love to eat
  - y: pizzas

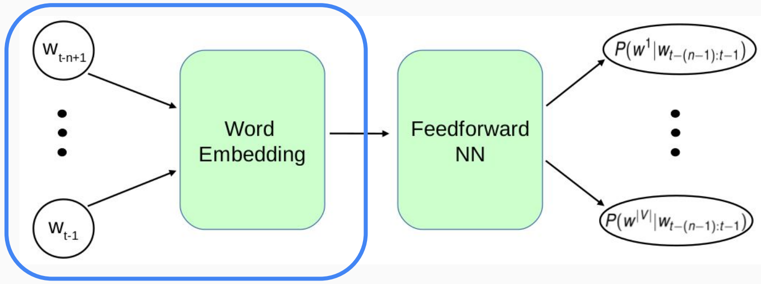


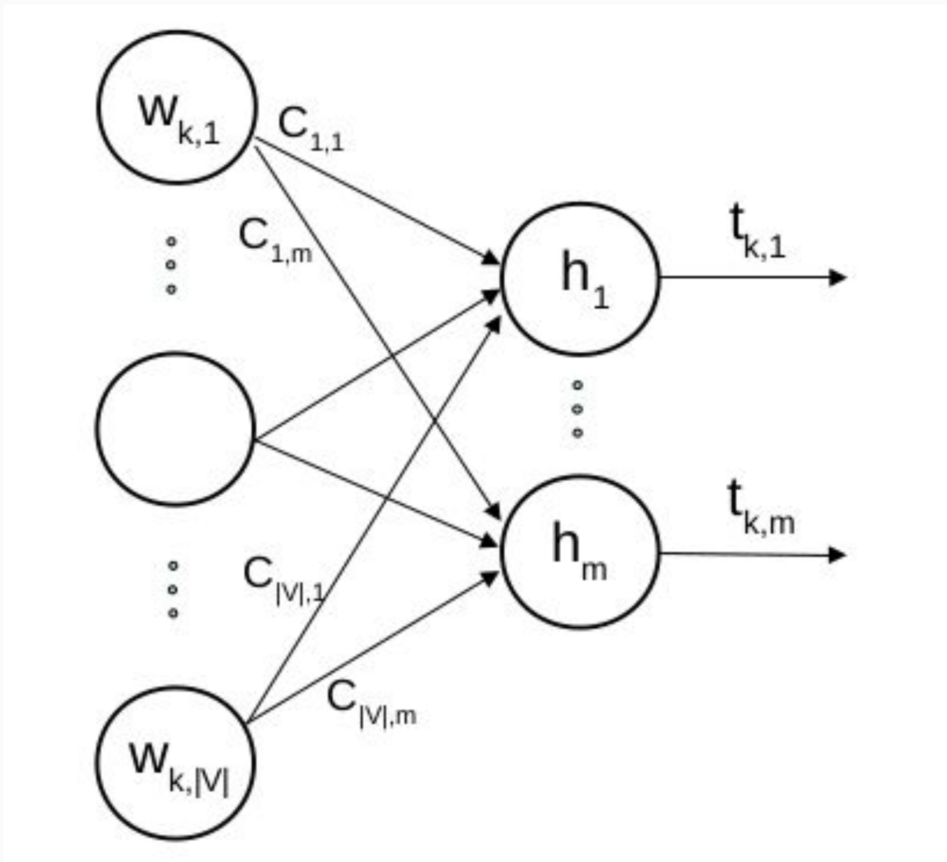
Bengio et al. (NIPS, 2000; JMLR, 2003)

- Instancia de test (x,y):
  - x: I love to eat
  - y: ?



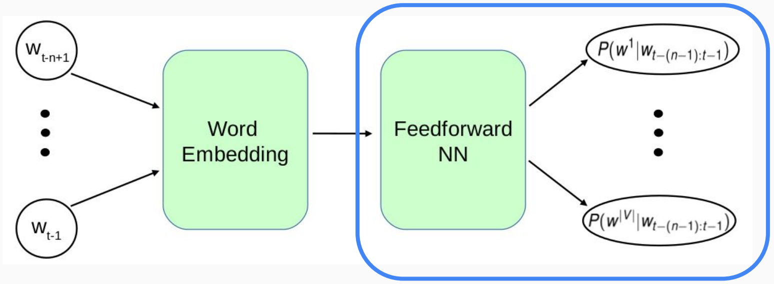
Bengio et al. (NIPS, 2000; JMLR, 2003)

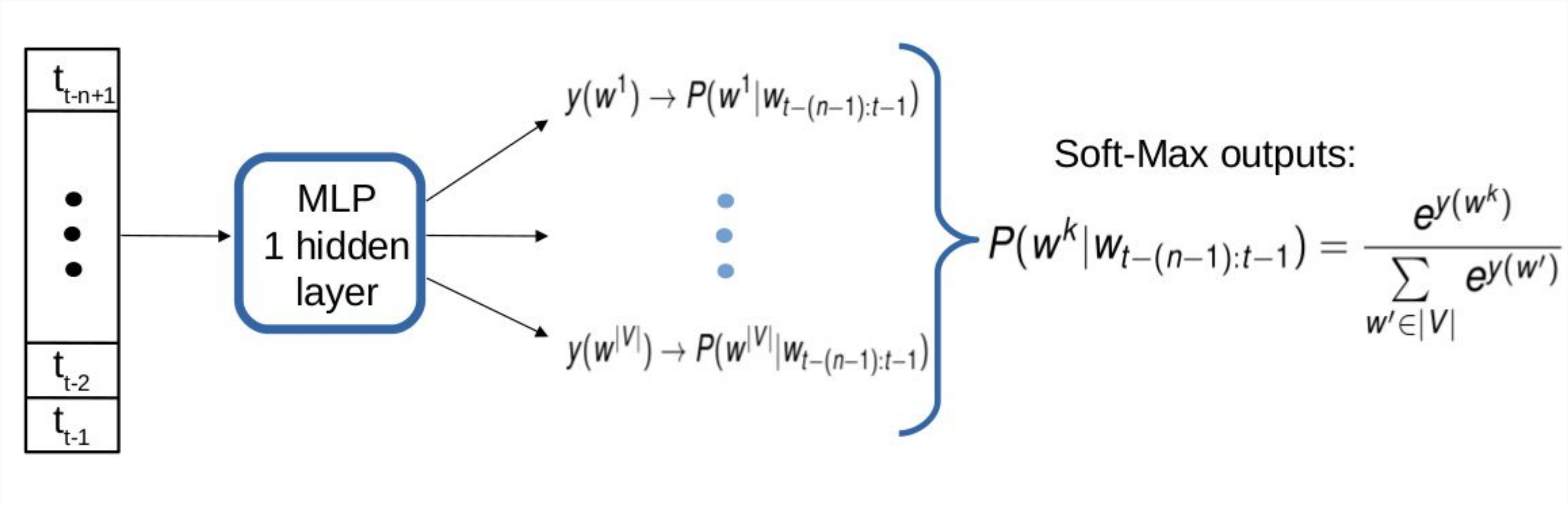




Aprender matriz real de proyección $C_{|V|*m}$

Bengio et al. (NIPS, 2000; JMLR, 2003)

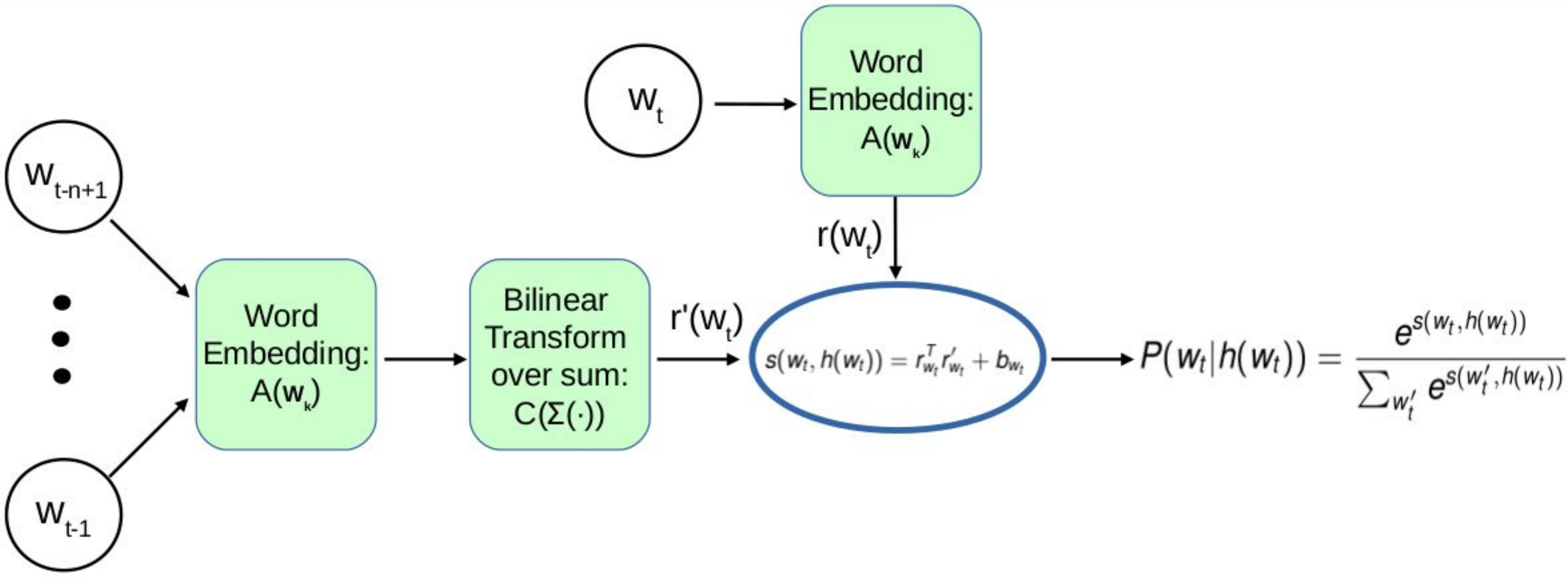




Aprender pesos para la feedforward NN

Otros NPLM...

Log-Bilinear Model (LBM) (Mnih & Hinton, 2007)



Otros NPLM...

Noise-Contrastive Estimation (NCE) + LBM extendido (Mnih & Teh, 2012)

- NCE: convertir problema de clasificación multinomial en problema de clasificación binaria
- Tiempo de entrenamiento significativamente más rápido (x14) que ML

NPLMs y Word Feature Learning

- NPLMs:
  - 1. *word-context probabilistic modeling*
  - (2. ***word-vector embedding***)
- Aprender representaciones vectoriales para alimentar varias aplicaciones de NLP

Mikolov et al. (2013)

- Nos fijamos en los *embeddings*
- Métodos anteriores son muy ineficientes para este fin:
  - × Muchos parámetros
  - × Costosos de entrenar
- Idea: simplificar modelos
  - ✓ Escalabilidad
  - ✓ Menos complejidad, más eficiencia
  - ✓ Mejores sistemas NLP

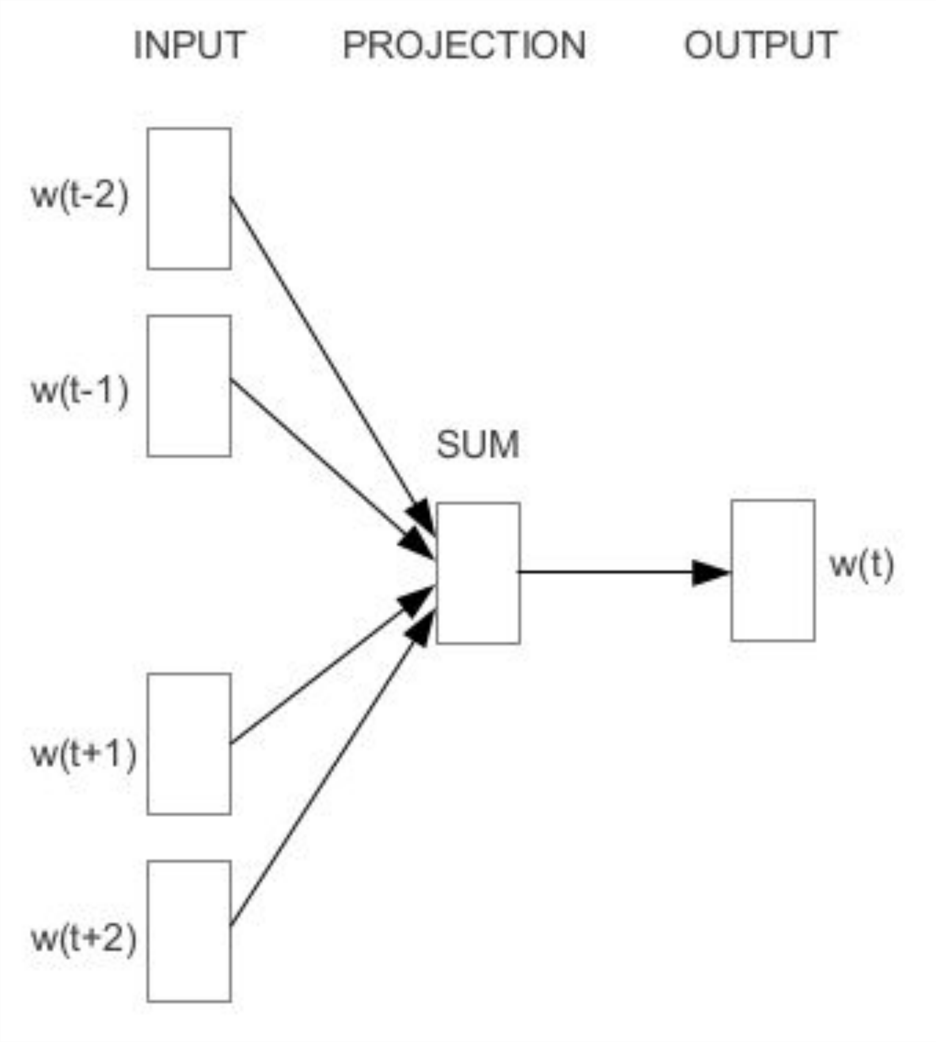
Mikolov et al. (2013)

- Simplificaciones:
  1. Remover capa oculta y directamente conectar *embeddings* con outputs del softmax. 2 modelos:
     - Continuous Bag-of-Words (CBoW)
     - Continuous Skip-gram (Skip-gram)
  2. Reemplazar soft-max por Hierarchical-softmax (HSMax), NCE o Negative Sampling (NS)

Mikolov et al. (2013)

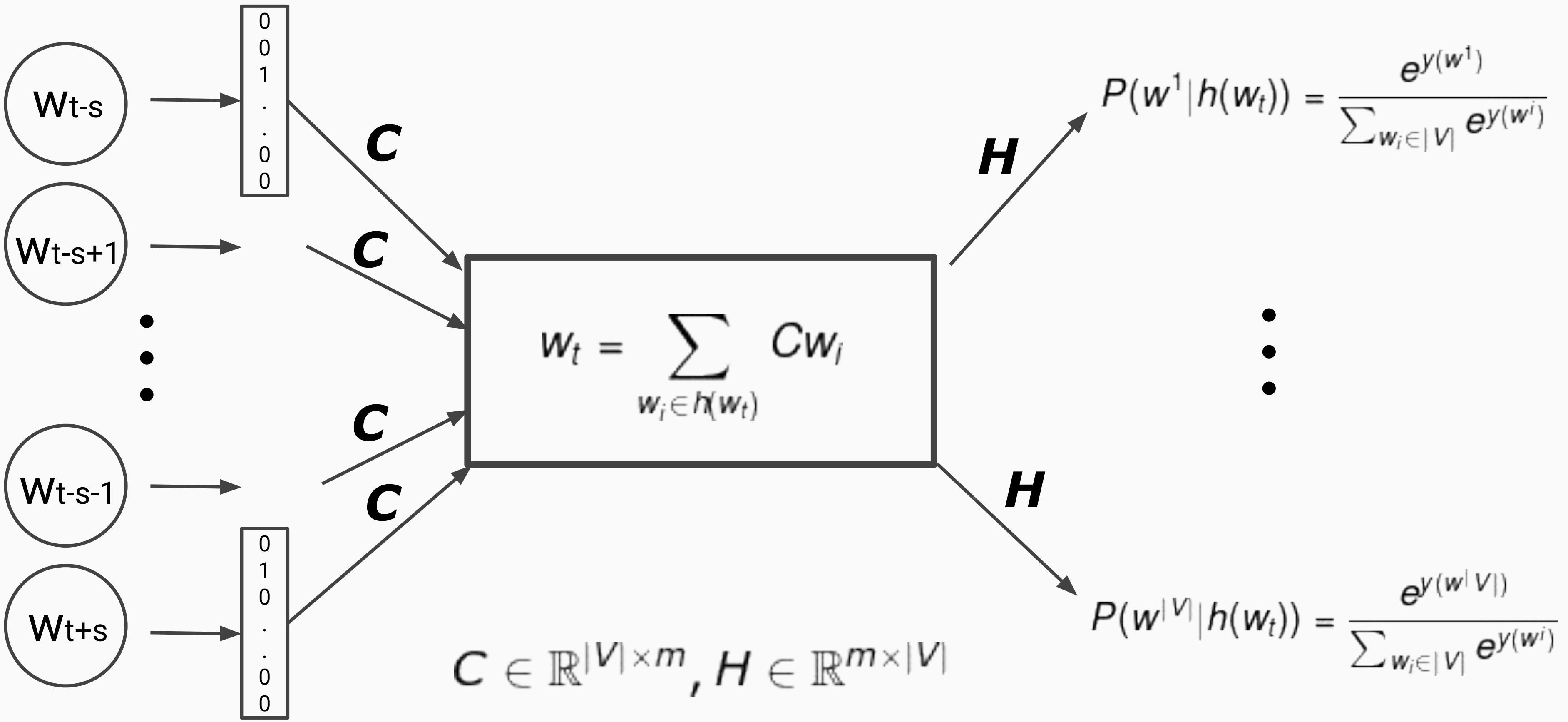
Continuous Bag-of-Words Model (CBoW)

- CBoW suma embeddings del contexto



Mikolov et al. (2013)

Continuous Bag-of-Words Model (CBoW)

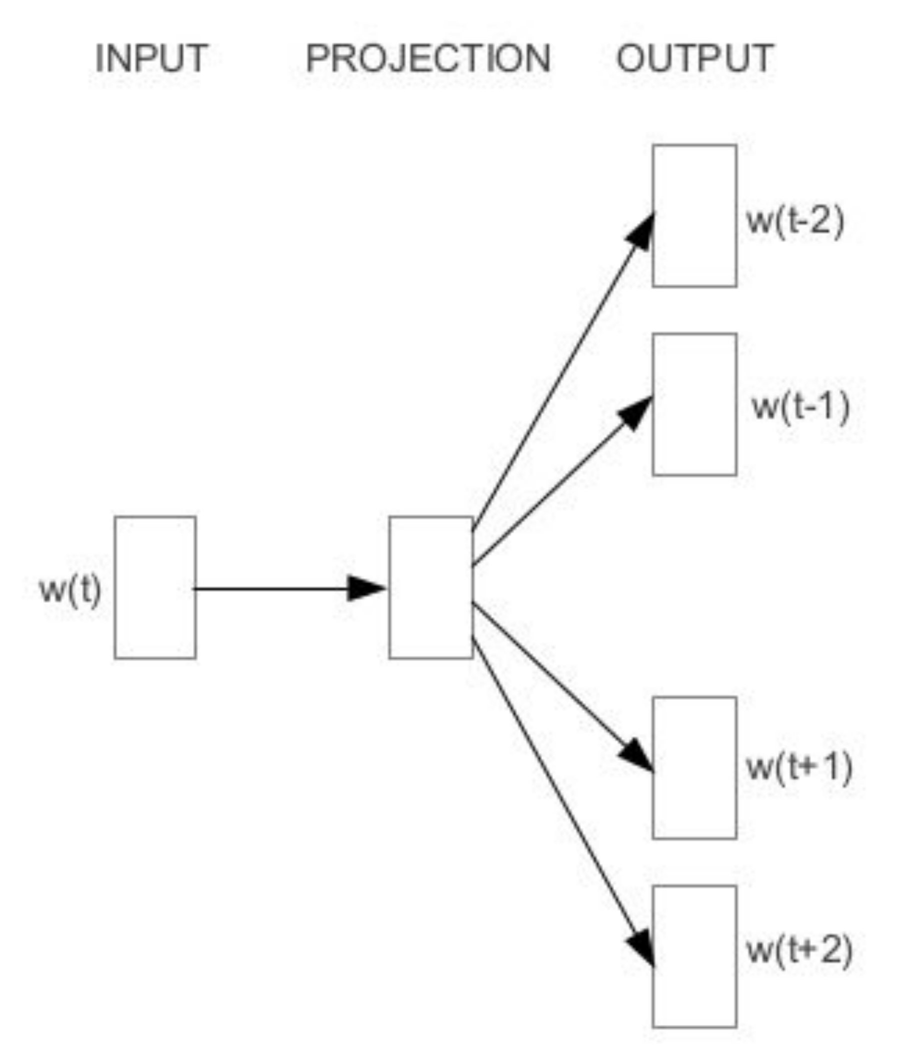


- Training: aprender matrices **C** y **H**

Mikolov et al. (2013)

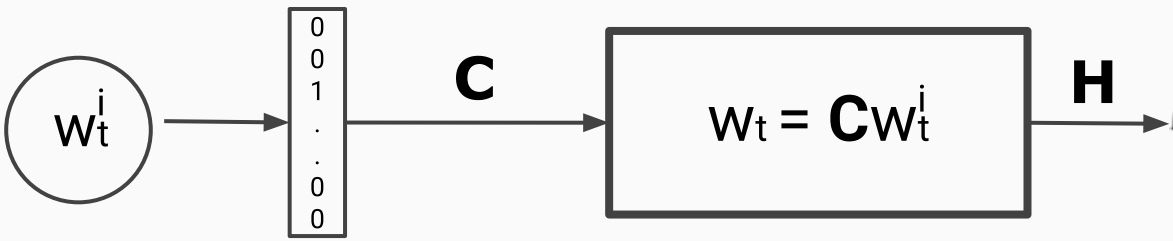
Continuous Skip-gram Model

- Skip-gram predice contexto a partir de una palabra central *w*



Mikolov et al. (2013)

Continuous Skip-gram Model



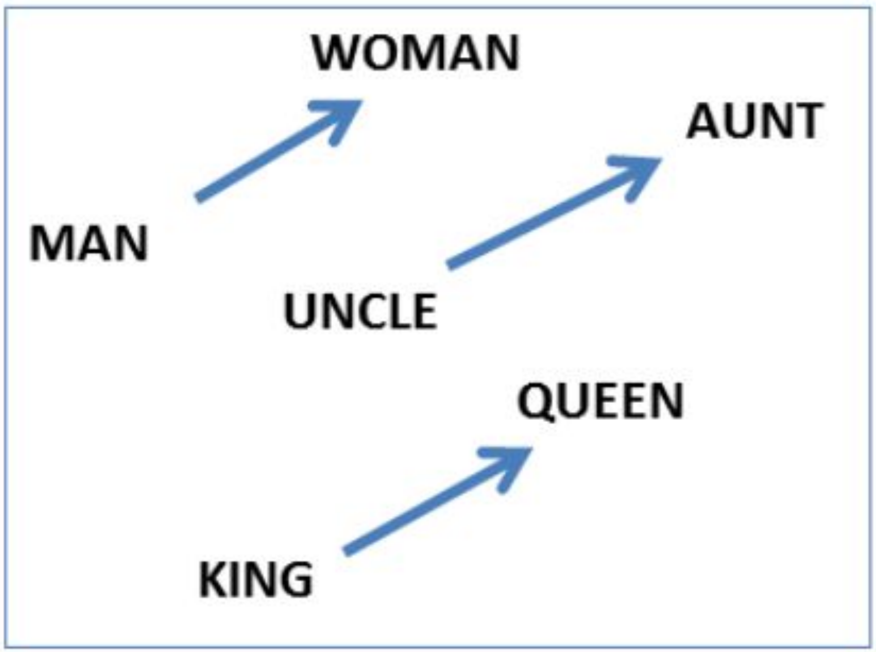
$$P(w^{k \in h(w_t)} | w_t) = \frac{e^{s(w_k, w_t)}}{\sum_{w'_k \in |V|} e^{s(w'_k, w_t)}}$$

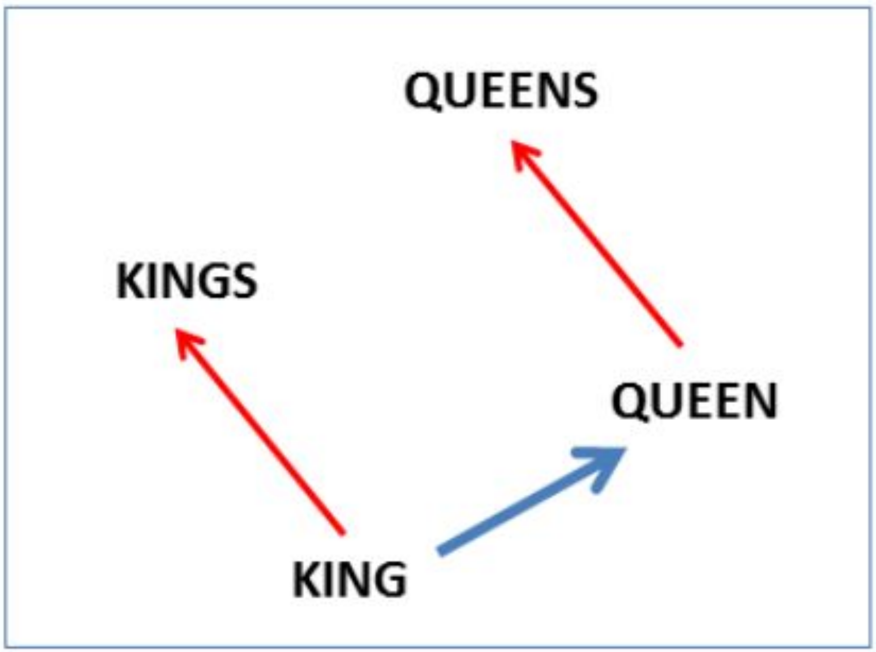
$$C \in \mathbb{R}^{|V| \times m}, H \in \mathbb{R}^{m \times |V|}$$

Mikolov et al. (2013)

Resultados

- Vector Offset Method
  - v(“King”) - v(“Man”) + v(“Woman”) ≈ v(“Queen”)





Mikolov et al. (2013)

Resultados

- Tareas sintácticas y semánticas: $y = x_b - x_a + x_c$
  - good:better bad:____
  - Germany:Berlin France:____

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Mikolov et al. (2013)

Resultados

- Training set: corpus de Google News. 1B~33B de palabras
- Vocabulario: 700K~1M
- Preguntas: 8K~10K de cada categoría

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Mikolov et al. (2013)

Resultados

Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

Mikolov et al. (2013)

Resultados

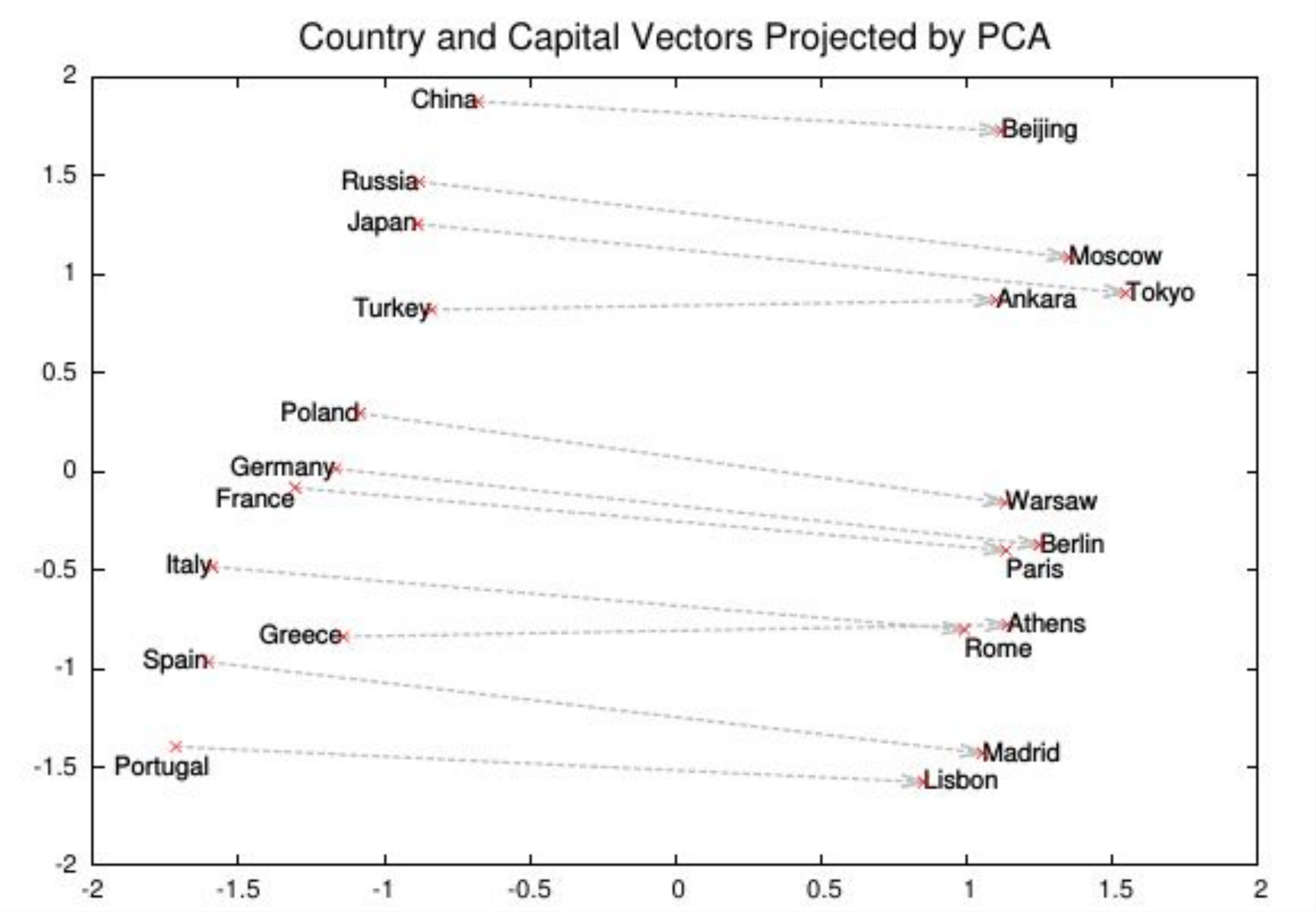
Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

- Additive Compositionality

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Mikolov et al. (2013)

Resultados



Word Embedding + Recommender Systems

Musto et al. (2015): *Word Embedding techniques for Content-based Recommender Systems: an empirical evaluation*

MovieLens	W2V		RI		LSI		U2U	I2I	BPRMF
Vector Size	300	500	300	500	300	500			
F1@5	**0.5056**	0.5054	0.4921	0.4910	0.4645	0.4715	**0.5217**	*0.5022*	0.5141
F1@10	**0.5757**	0.5751	0.5622	0.5613	0.5393	0.5469	**0.5969**	0.5836	0.5928
F1@15	0.5672	**0.5674**	0.5349	0.5352	0.5187	0.5254	**0.5911**	0.5814	0.5876
DBbook	W2V		RI		LSI		U2U	I2I	BPRMF
	300	500	300	500	300	500			
F1@5	0.5183	**0.5186**	0.5064	0.5039	0.5056	0.5076	0.5193	*0.5111*	**0.5290**
F1@10	0.6207	0.6209	0.6239	0.6244	0.6256	**0.6260**	*0.6229*	*0.6194*	**0.6263**
F1@15	0.5829	0.5828	0.5892	0.5887	0.5908	**0.5909**	*0.5777*	*0.5776*	*0.5778*

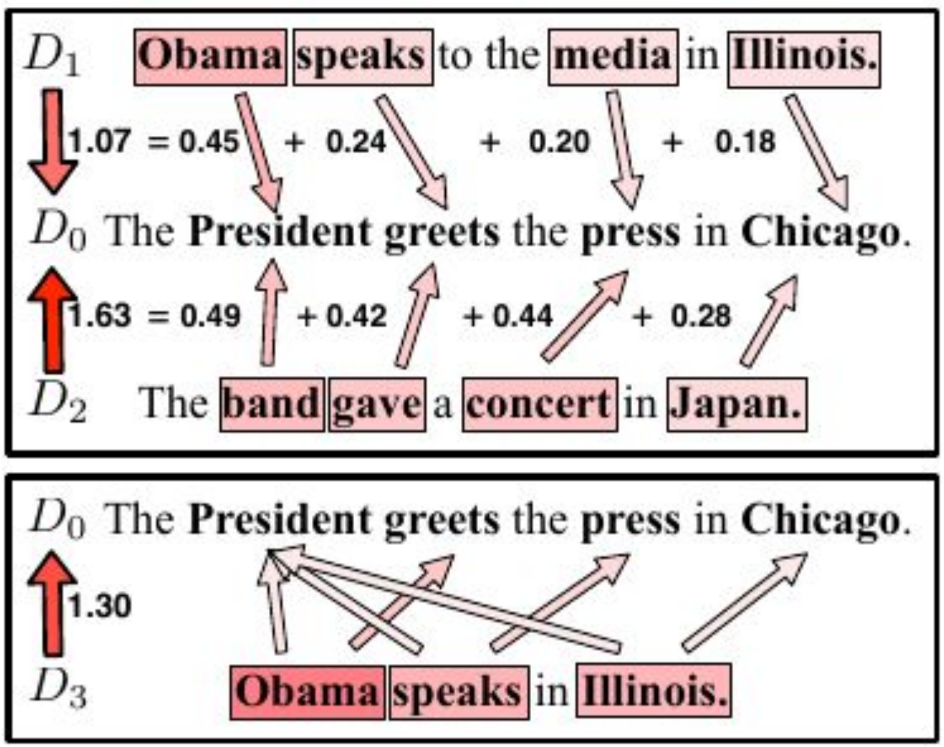
Gulcin et al. (2016): *From Word Embeddings to Item Recommendation*

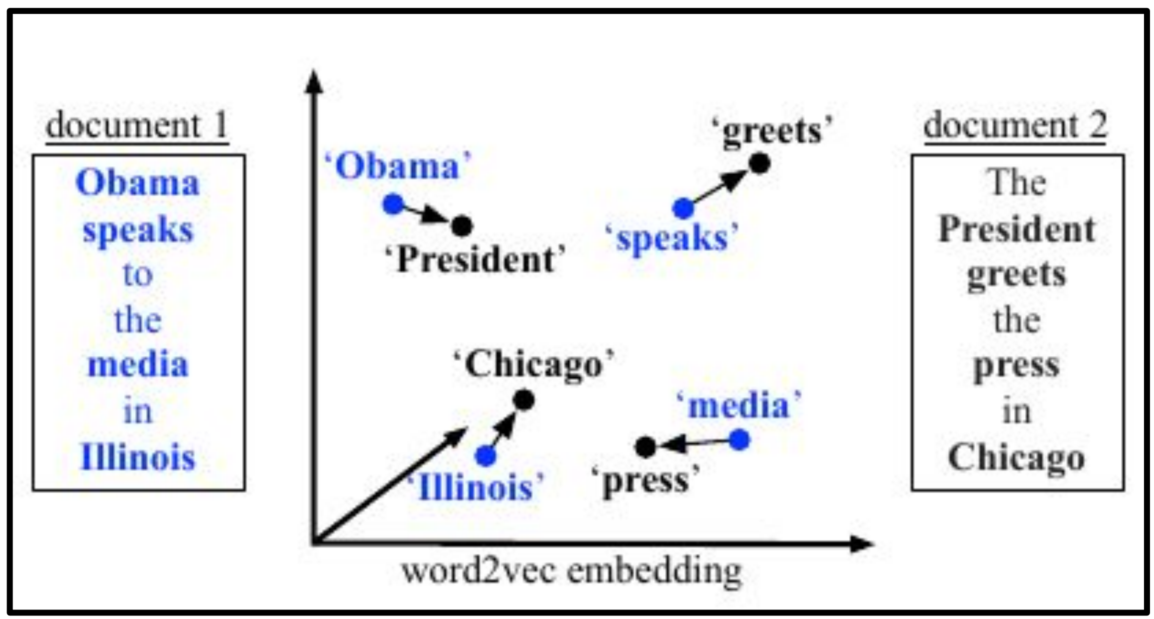
- Recomendación de lugares usando *features* no-textuales (check-ins)

Word Mover's Distance (WMD)

Kusner et al. (2015): ***From Word Embeddings To Document Distances*.**

Distancia mínima que las palabras embebidas de un documento tienen que "viajar" para llegar a las palabras embebidas del otro documento





Referencias

- Bengio et al. (2003): *A Neural Probabilistic Language Model.*
- Mikolov et al. (Jun/2013): *Linguistic Regularities in Continuous Space Word Representations.*
- Mikolov et al. (Sep/2013): *Efficient Estimation of Word Representations in Vector Space.*
- Mikolov et al. (Oct/2013): *Distributed Representations of Words and Phrases and their Compositionality.*
- Musto et al. (2015): *Word Embedding techniques for Content-based Recommender Systems: an empirical evaluation.*
- Gulcin et al. (2016): *From Word Embeddings to Item Recommendation.*
- Kusner et al. (2015): *From Word Embeddings To Document Distances*.

Links de interés

- Intuitive explanation of Noise Contrastive Estimation (NCE) loss?
- Candidate Sampling
- The amazing power of word vectors