Calculating Similarities among Languages using Word2Vec

Presented by: Yizhong Wang

Presentation date: 2015.11.25

Paper Info

Title:	Exploiting Similarities among Languages for Machine Translation
Authors:	**Tomas Mikolov**, Quoc V. Le, Ilya Sutskever
Date:	17 Sep 2013
Cited by:	103
Goal:	Find translation pairs between languages automatically
Idea:	Learn word vectors for each language based on monolingual data and lean a linear mapping between vector spaces of languages

Outline

- **Basics of Word2Vec**
  - Representation of word
  - Magic of Word2Vec
  - Skip-gram and CBOW model
  - Training and efficiency
- **Relationship between Languages**
  - Visualization of words in English and Spanish
  - Formalizing the linear relationship
  - Training and predicting
- **Experiments and Results**
  - Baseline techniques
  - Training and testing data
  - Results and conclusion

Basics of Word2Vec

Representation of word

1. WordNet： hypernyms (red-color), synonyms(like-enjoy).
2. One-hot vector:

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

3. Cooccurrence matrix:

I like NLP.
I like soccer.

Count	I	like	NLP	soccer
I	0	2	0	0
like	2	0	1	1
NLP	0	1	0	0
soccer	0	1	0	0

4. Dense vector: low dimensional vector:

SVD
Word2Vec
Deep Learning

Quiz 1:

Open question:

What weakness do you think cooccurrence matrix approach might have?

I like NLP.
I like soccer.

Count	I	like	NLP	soccer
I	0	2	0	0
like	2	0	1	1
NLP	0	1	0	0
soccer	0	1	0	0

Magic of Word2Vec

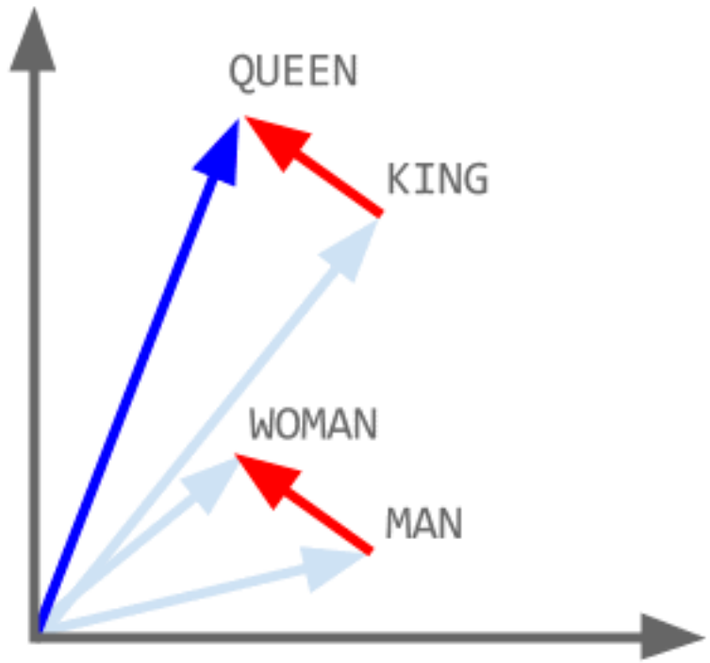
- A class of **two-layer** neural network models
- Input: unlabelled training corpus
- Output: vector v_{wi} for each word w_i in vocabulary V

$$dim(v_{w_i}) \ll |V|$$

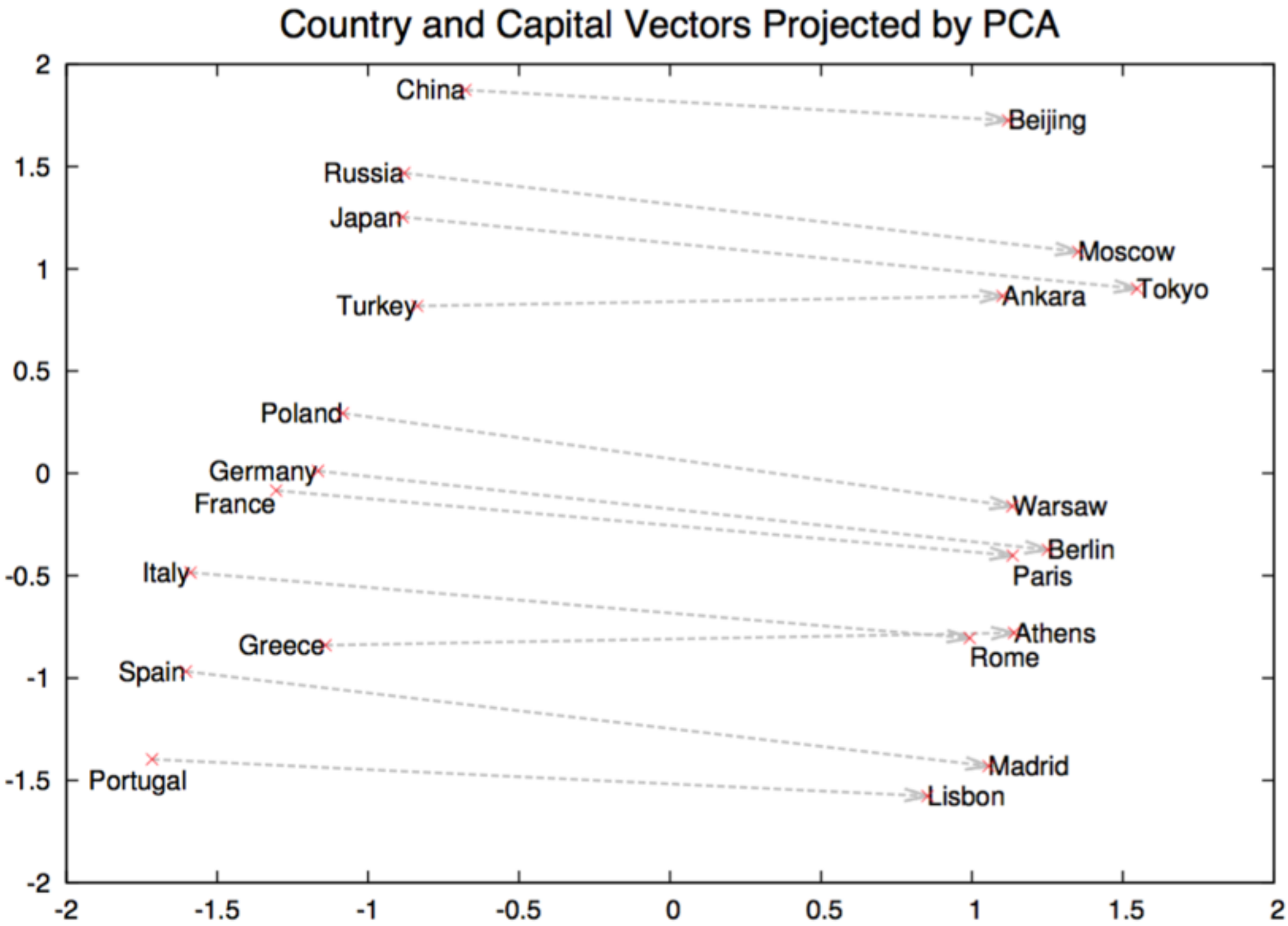
- Surprising property: **linear regularity**

$$v_{king} - v_{man} + v_{woman} = v_{queen}$$

$$v_{books} - v_{book} = v_{phones} - v_{phone}$$

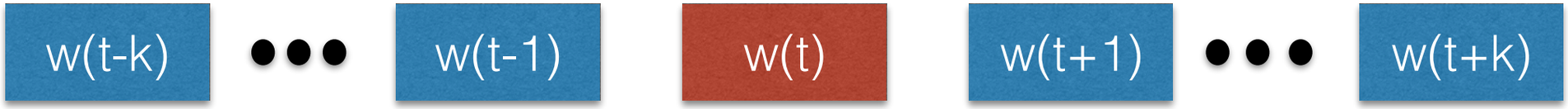


Magic of Word2Vec



Skip-gram and CBOW model

- Given a central word and its context with window size k



- CBOW (Continuous Bag-of-Words) model:

Predict the center: $$\max \frac{1}{T}\sum_{t=1}^{T}[\sum_{j=-k}^{k} \log p(w_t|w_{t+j})]$$

- Skip-gram model:

Predict the context: $$\max \frac{1}{T}\sum_{t=1}^{T}[\sum_{j=-k}^{k} \log p(w_{t+j}|w_t)]$$

Quiz 2:

Tell the names of the two models in Word2Vec and what is their main difference.

Skip-gram model

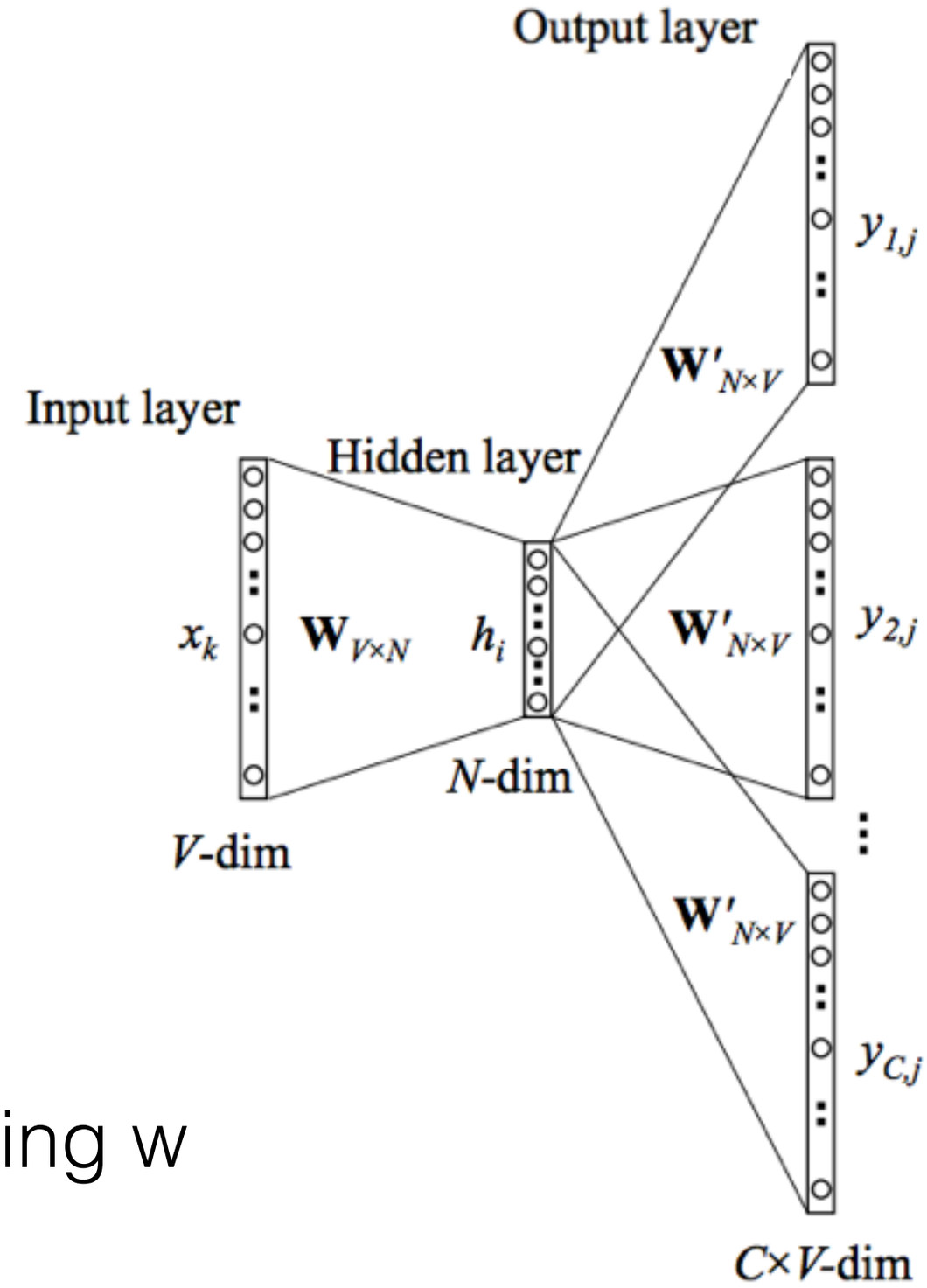
$$\max \frac{1}{T} \sum_{t=1}^{T} [\sum_{j=-k}^{k} \log p(w_{t+j}|w_t)]$$

$$p(w_i|w_j) = \frac{\exp\left({u_{w_i}}^\top v_{w_j}\right)}{\sum_{l=1}^{V} \exp\left({u_l}^\top v_{w_j}\right)}$$

u_w : input vector of w

v_w : output vector of w

$u_w + v_w$: final vector representing w



Training and efficiency

- Training techniques:

  Stochastic Gradient Descent + Backpropagation

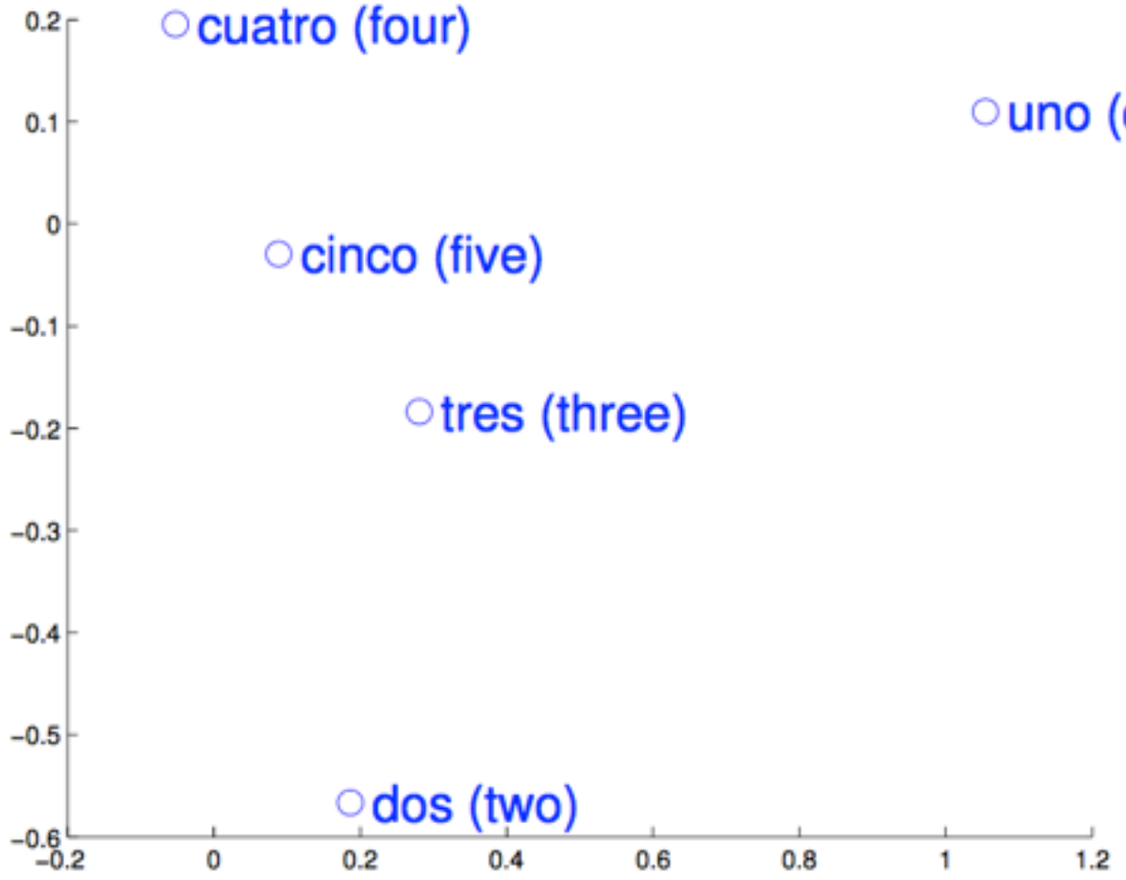
- Training efficiency:

  Optimized single-machine implementation can train on >100 billion words in one day

Relationship between languages

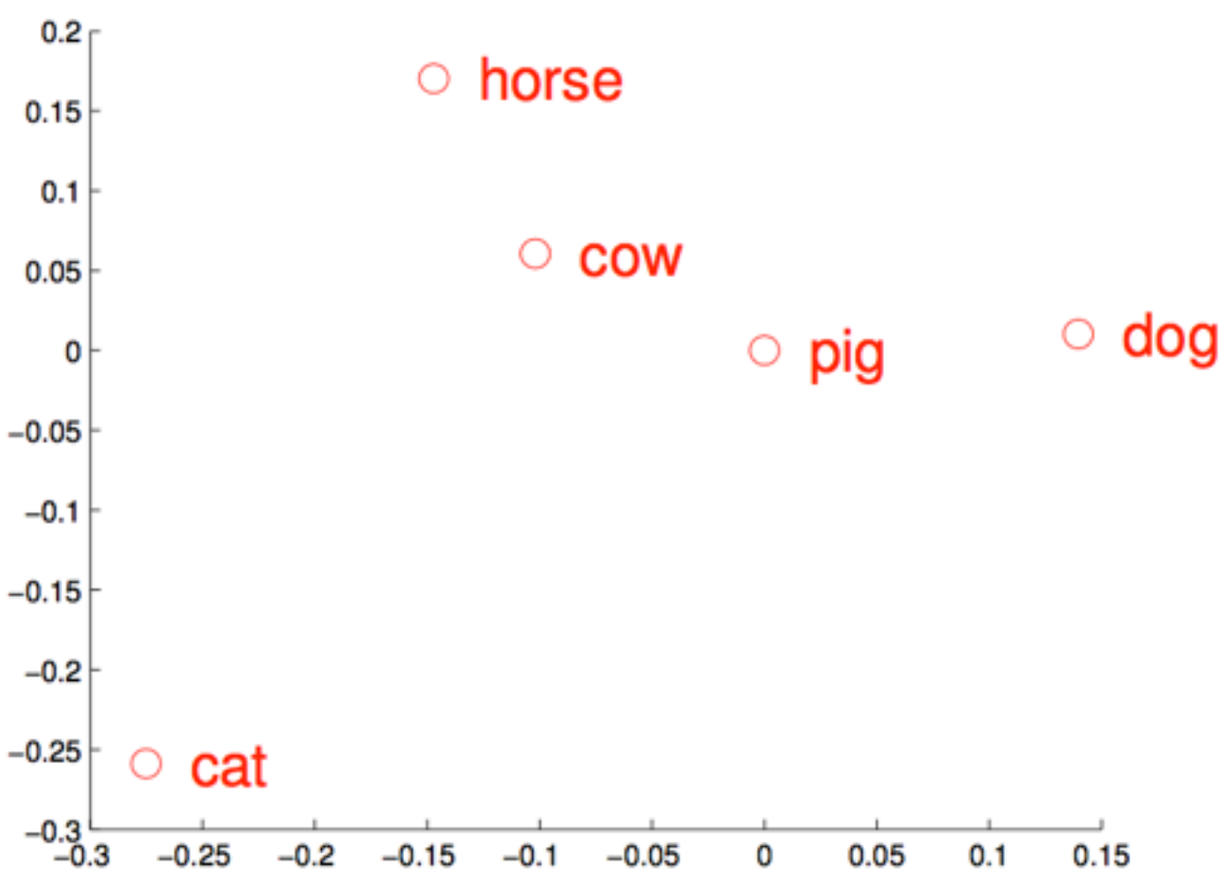
Visualization of words in English and Spanish

Numbers in English

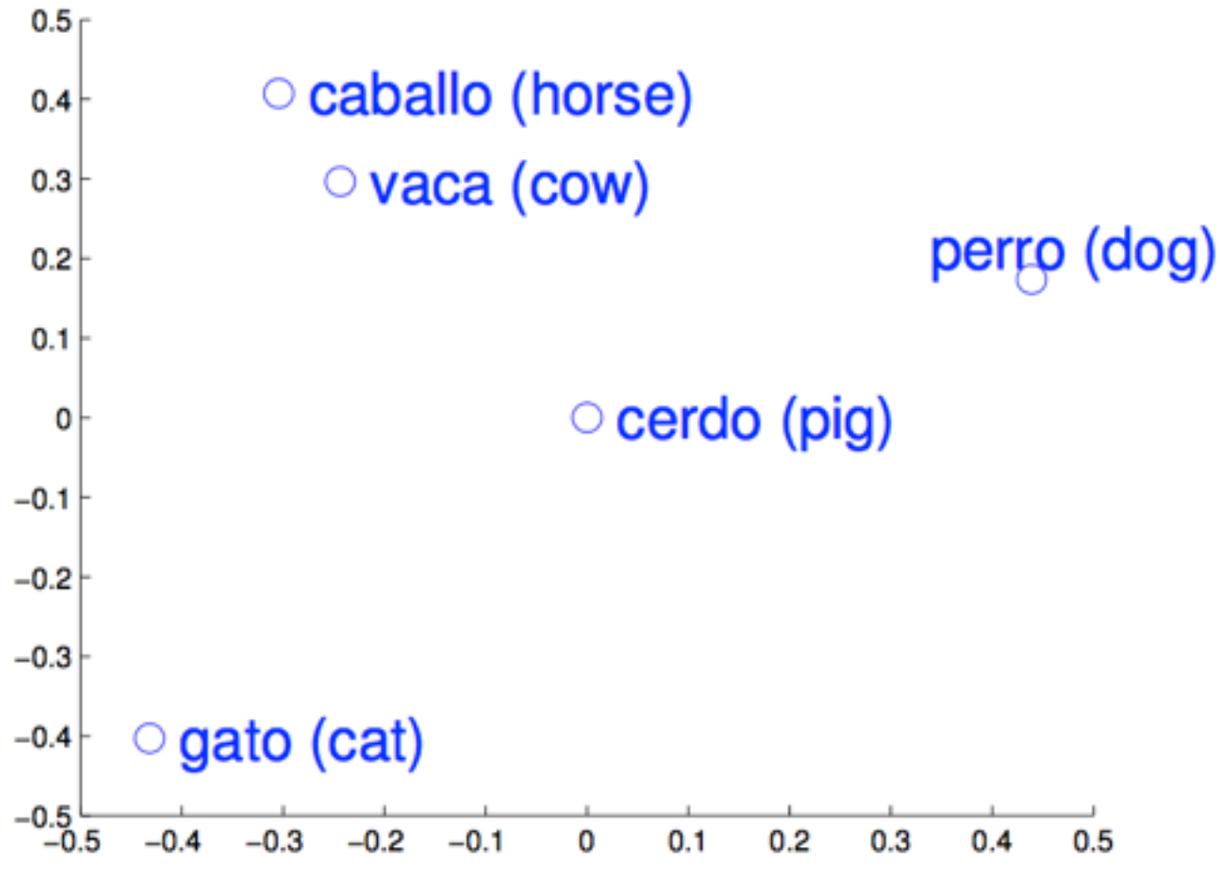


Numbers in Spanish

Visualization of words in English and Spanish



Animals in English



Animals in Spanish

Formalizing the relationship

Corresponding pairs have similar geometric arrangement

↓

The relationship between vector spaces

can be captured by rotation and scaling

↓

linear transformation can be represented by matrix W

↓

Our goal is to calculate W that:

$$argmin_W \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

Training and predicting

- To collect the training and testing pairs:

Google Translation

- To train the transformation matrix W:

Stochastic Gradient Descent

- To predict the corresponding Spanish word z_i for English word x_i

Find the word closest to Wx_i in the Spanish space

Experiments and Results

Training and testing data

1. Monolingual datasets from WMT11 for word vector training

Language	Training tokens	Vocabulary size
English	575M	127K
Spanish	84M	107K
Czech	155M	505K

2. Google translated pairs:

6K most popular words
- 5K for training matrix W
- 1K for testing

Baseline techniques to find pairs

1. Editing distance between words in different languages

2. Using similarity of word concurrence vector

- Bilingual Dictionary: I-我, love-喜欢, and-和, basketball-篮球
- Count occurrence of in-dictionary words for each test word.
- The cooccurrence vector is mapped to the other language using dictionary.

Results and conclusion

Accuracy of word translation methods

Translation	Edit Distance		Word Co-occurrence		Translation Matrix		ED + TM	
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
En → Sp	13%	24%	19%	30%	33%	51%	43%	60%
Sp → En	18%	27%	20%	30%	35%	52%	44%	62%
En → Cz	5%	9%	9%	17%	27%	47%	29%	50%
Cz → En	7%	11%	11%	20%	23%	42%	25%	45%

P@1 : percentage that top 1 closest word matches the target

P@5 : percentage that top 5 closest words includes the target

Results and conclusion

Examples of translations

Spanish word	Computed English Translations	Dictionary Entry
emociones	emotions emotion feelings	emotions
protegida	wetland undevelopable protected	protected
imperio	dictatorship imperialism tyranny	empire
determinante	crucial key important	determinant
preparada	prepared ready prepare	prepared

Examples of error detection

English word	Computed Czech Translation	Dictionary Entry
said	řekl *(said)*	uvedený *(listed)*
will	může *(can)*	vůle *(testament)*
did	udělal *(did)*	ano *(yes)*
hit	zasáhl *(hit)*	hit -
must	musí *(must)*	mošt *(cider)*
current	stávající *(current)*	proud *(stream)*
shot	vystřelil *(shot)*	shot -

Thanks,

Q & A