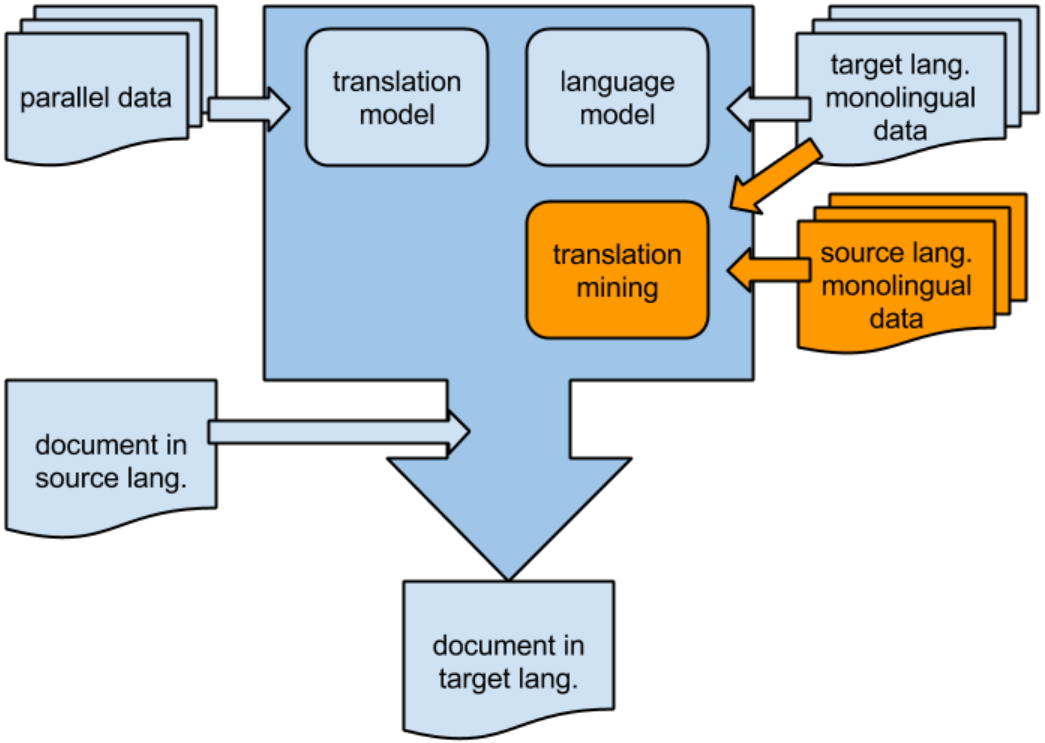
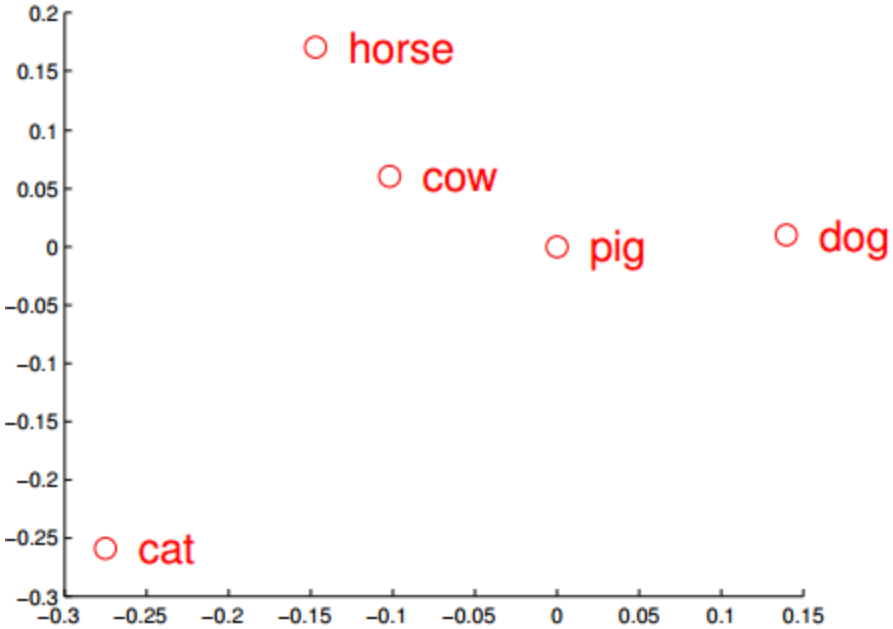
Exploiting Similarities among Languages for Machine Translation

(Mikolov et al 2013)


parallel data
translation model
language model
target lang. monolingual data
translation mining
source lang. monolingual data
document in source lang.
document in target lang.

Intuition



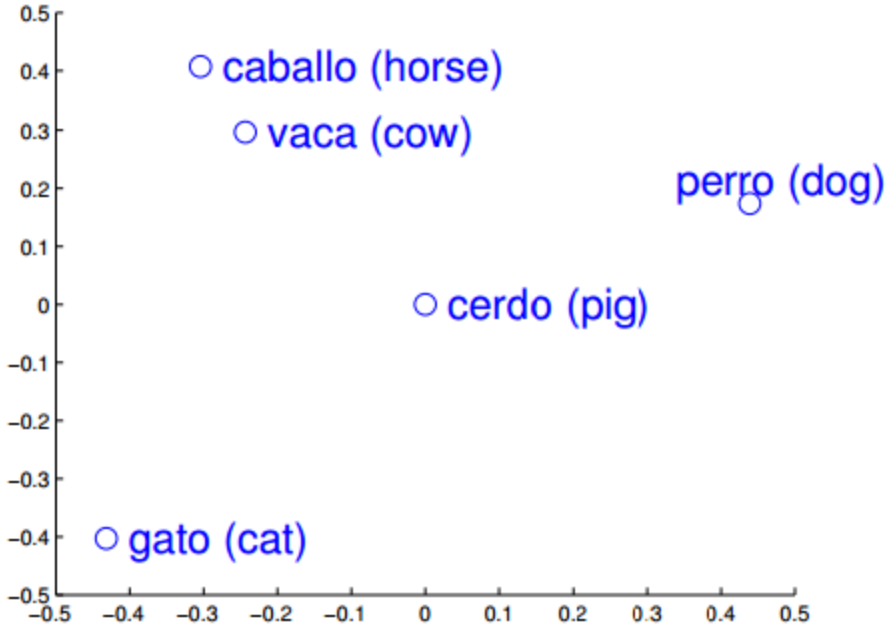


Image by Mikolov et al.

How it works

Acquire:

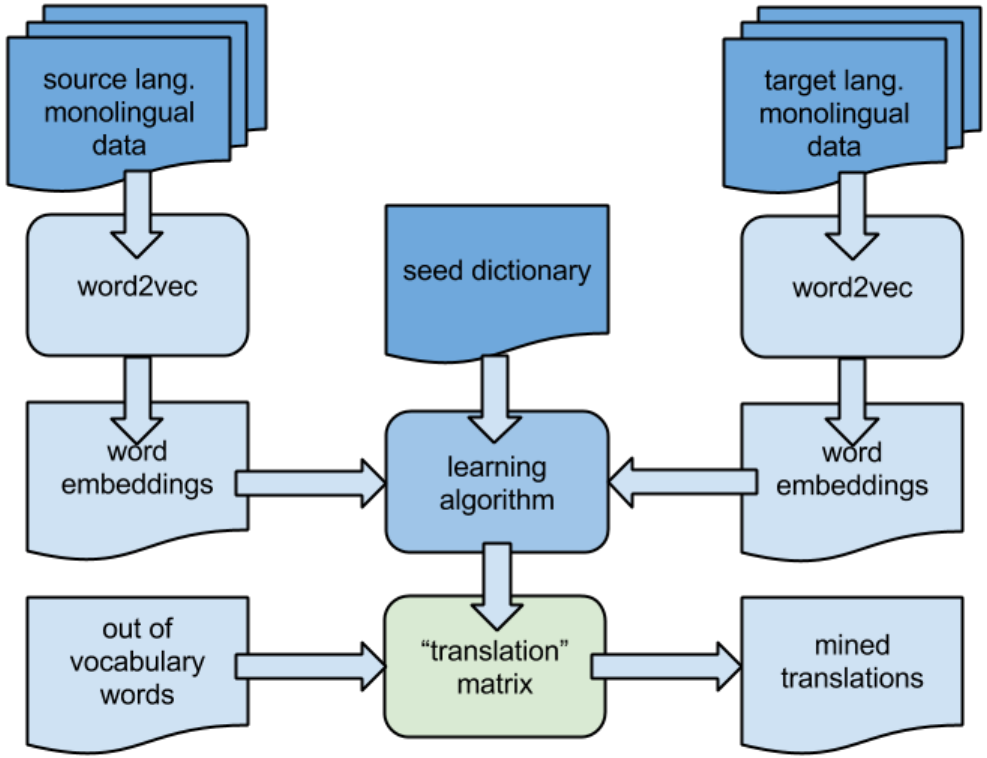
Monolingual data (hundreds of millions of words)

Seed dictionary (hundreds to thousands of words)

Learn:

Translation matrix (SGD or other learning algorithm)

Distributed representations of words (CBOW or Skip-gram models as implemented in word2vec)


source lang. monolingual data
target lang. monolingual data
word2vec
seed dictionary
word2vec
word embeddings
learning algorithm
word embeddings
out of vocabulary words
“translation” matrix
mined translations

Distributed Word Representations

COBW - word predicted from its context

Skip-gram - context predicted from the word

- Both give representations of a similar quality
- Both are proposed by T. Mikolov, any other word representations might be used instead
- Implemented in word2vec toolkit

Translation Matrix

Let x, y be distributed word representations for two words in a translation pair.

Then we want to learn a matrix W such that:

$$y = Wx$$

This gives an optimization problem over translation pairs in the seed dictionary:

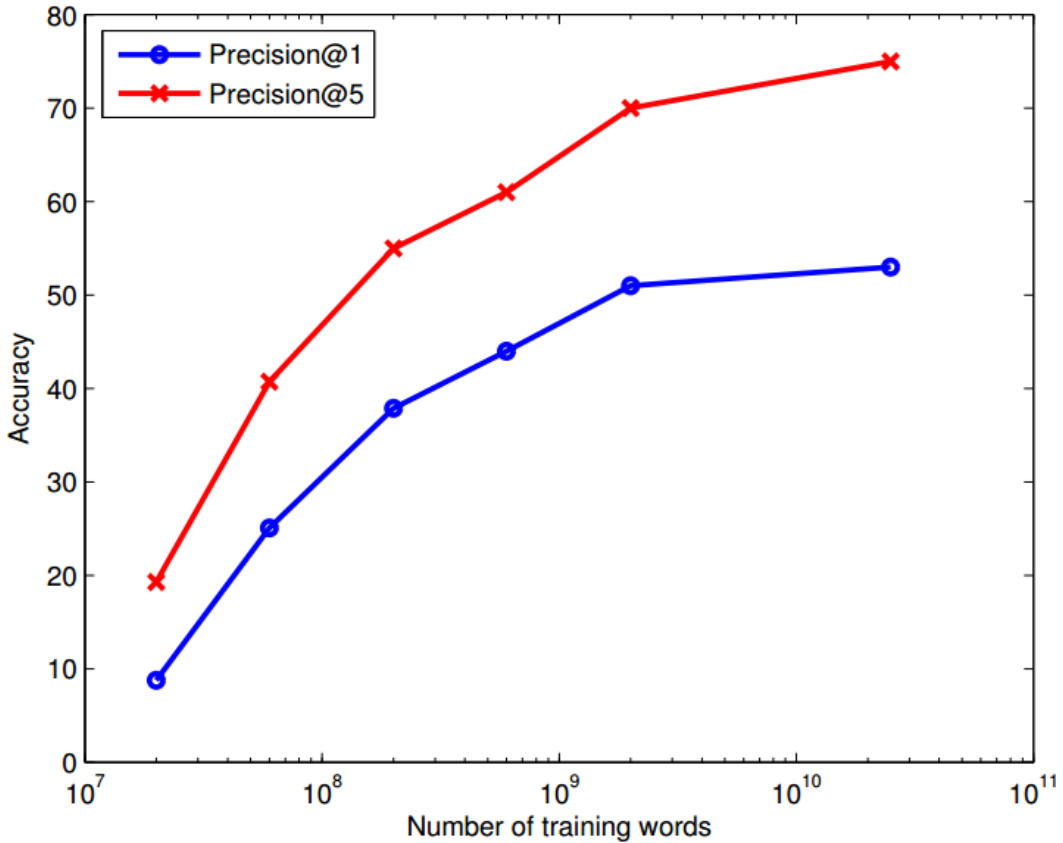
$$\min_W \sum || Wx_i - y_i||^2$$

Results

Translation	Edit Distance		Word Co-occurrence		Translation Matrix	
	P@1	P@5	P@1	P@5	P@1	P@5
En → Sp	13%	24%	19%	30%	33%	51%
Sp → En	18%	27%	20%	30%	35%	52%
En → Cz	5%	9%	9%	17%	27%	47%
Cz → En	7%	11%	11%	20%	23%	42%

Trained on WMT11 datasets (575M English tokens, 84M Spanish tokens, 155M Czech tokens)

With more data…



Performance doubles if the amount of data increases by two orders of magnitude.

Precision at 1 and 5 as the size of monolingual training sets increase. (En to Sp)

English word	Computed Spanish Translation	Dictionary Entry
pets	mascotas	mascotas
mines	minas	minas
unacceptable	inaceptable	inaceptable
prayers	oraciones	rezo
shortstop	shortstop	campocorto
interaction	interacción	interacción
ultra	ultra	muy
beneficial	beneficioso	beneficioso
beds	camas	camas
connectivity	conectividad	conectividad
transform	transformar	transformar
motivation	motivación	motivación

Examples by Mikolov et al

Questions or Comments?