Accurate Cross-lingual Projection between Count-based Word Vectors by Exploiting Translatable Context Pairs

Shonosuke Ishiwatari♣, Nobuhiro Kaji♢♡, Naoki Yoshinaga♢♡, Masashi Toyoda♢, Masaru Kitsuregawa♢♠

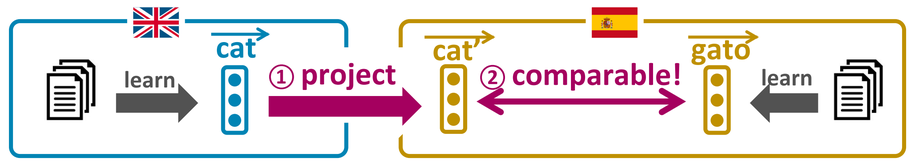
{ishiwatari, kaji, ynaga, toyoda, kitsure} @tkl.iis.u-tokyo.ac.jp

♣The University of Tokyo, ♢IIS, the University of Tokyo, ♡NICT, Japan, ♠NII, Japan

Overview

Problem: Word vectors in different languages are not comparable because they are learned from different corpora

Approach: Project a vector from a language into another language space [Fung+, 98][Mikolov+, 13]

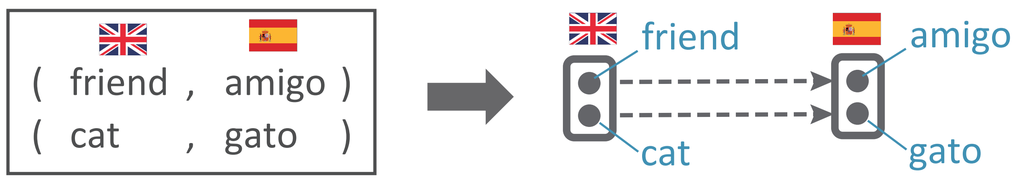


Idea: Incorporate previous approaches to realize an accurate projection between word vectors!

Previous studies

Dictionary-based approach [Fung+, 98]

Directly map count-based word vectors for each dimension by using a dictionary

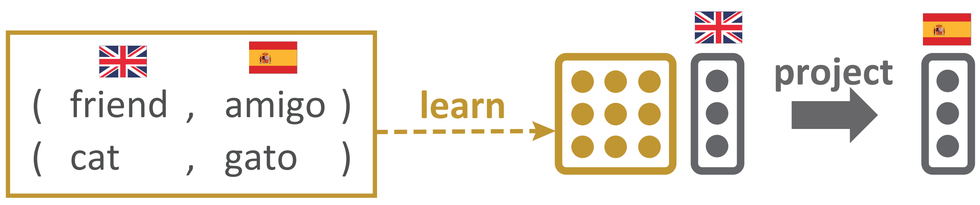


☺ Can use relationships between context words

☹ Only one-to-one mapping is allowed

Learning-based approach [Mikolov+, 13]

Learn a linear transformation (i.e., a matrix) between predict-based word vectors

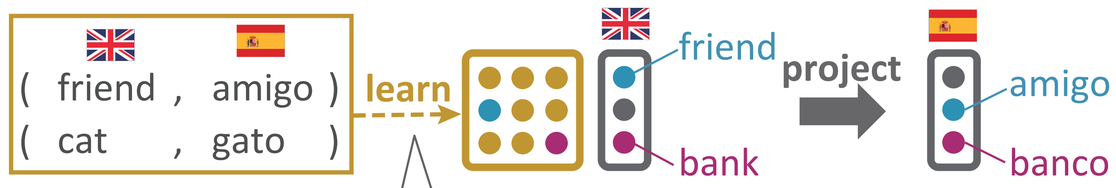


☺ Can find many-to-many correlation between the elements automatically

☹ Can't utilize the context words to learn a model

Proposed approach

- Use count-based word vectors to utilize the knowledge about context words
- The existing knowledge are obtained from
  - **training set** and
  - **surface similarity** between context words
- Weight the corresponding elements in the matrix with two bonus terms



Objective function:

$$\min_W \sum_{i=1}^{n} \| Wx_i - z_i \|^2 - \beta_{train} \sum_{(j,k)\in\mathcal{D}_{train}} w_{jk} - \beta_{sim} \sum_{(j,k)\in\mathcal{D}_{sim}} w_{jk}$$

[Mikolov+, 13]

x: source vector, z: target vector, W: matrix

☺ Can find many-to-many correlation between the elements automatically

☺ Can use relationships between context words

Experiments

Data:

- Wikpedia dumps for learning word vectors
- Open Multilingual Wordnet for extracting bilingual pairs to train and test the projection
  - Most frequent 11k words for train
  - The subsequent 1k words for test

Evaluation procedure:

1. Given a word vector in the source language
2. Translate the vector into the target language
3. Choose the top-n (n = 1, 5) similar vectors in the target language
4. Check if the correct translation is included in the n vectors

The accuracy of the translation

Testset	[Mikolov+, 13] P@1	[Mikolov+, 13] P@5	[Fung+, 98] P@1	[Fung+, 98] P@5	Proposed P@1	Proposed P@5
Es → En	7.5%	22.0%	45.7%	61.1%	**54.7%**	**67.6%**
En → Es	7.1%	18.9%	11.9%	26.1%	**31.3%**	**49.6%**
Jp → Cn	5.4%	13.8%	9.3%	22.2%	**15.5%**	**34.0%**
Cn → Jp	2.9%	11.3%	11.6%	26.8%	**13.1%**	**27.9%**
En → Jp	4.9%	13.3%	5.4%	13.9%	**19.3%**	**37.1%**
Jp → En	6.5%	19.1%	22.3%	37.4%	**32.5%**	**51.9%**

Top 5 translation candidates of "ニワトリ" (chicken) (Jp → En)

[Mikolov+, 13]	[Fung+, 98]	Proposed
1. kind	1. animal	**1. chicken**
2. call	2. rabbit	**2. animal**
3. frequently	3. eat	**3. rabbit**
4. turn	4. chicken	**4. eat**
5. make	5. fish	**5. wild**
⋮		
844. chicken		

correct answer

Impact of the size of training data (Es → En)