Cross-lingual word embedding

Sommaire

Introduction

Based on “A Survey of Cross-lingual Word Embedding Model” by Ruder, Vulic and Sogaard

Represent vocabulary of 2 or more languages in one common vector space

Used to :

- Improve monolingual similarity
- Support cross-lingual transfer

Referencies

http://ruder.io/cross-lingual-embeddings/index.html

https://arxiv.org/pdf/1706.04902.pdf

I Type of data

Mostly on bilingual signal

--> How to select resources ?

--> Type of alignment

--> Comparability

Alignment

Word alignment :

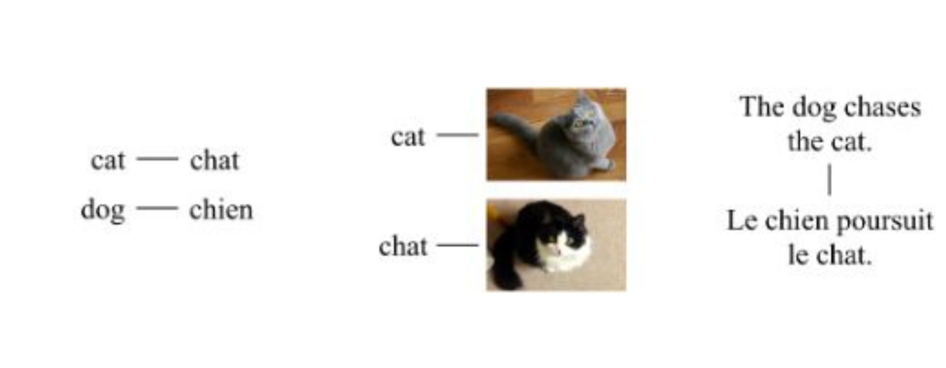
Pairs of translation between words

Sentence alignment :

Mainly use for MT

Document alignment :

Wikipedia



II Data everywhere

What do we want ?

What are available data ?

What is the purpose of the corpora ?

How to create or collect our own data in order to work on a targeted subject ?

Games with purposes - Books - Writings

A word/sentence parallel corpora :

-> translating game word by word (Dualingo) or sentence by sentence based on the bleu metric

A word/sentence comparable corpora :

-> image or idea to word or sentence

A document comparable corpora :

-> dissertation or book already translated (Gutenberg project)

III Word level alignment model

- Mapping based approaches
    - Minimizing mean squared error
    - CCA based mapping
- Word level based on pseudo Bilingual corpus
- Joint models
    - Bilingual language model
- Word level alignment methods with comparable data
    - POS tag equivalence
    - Grounding language in image

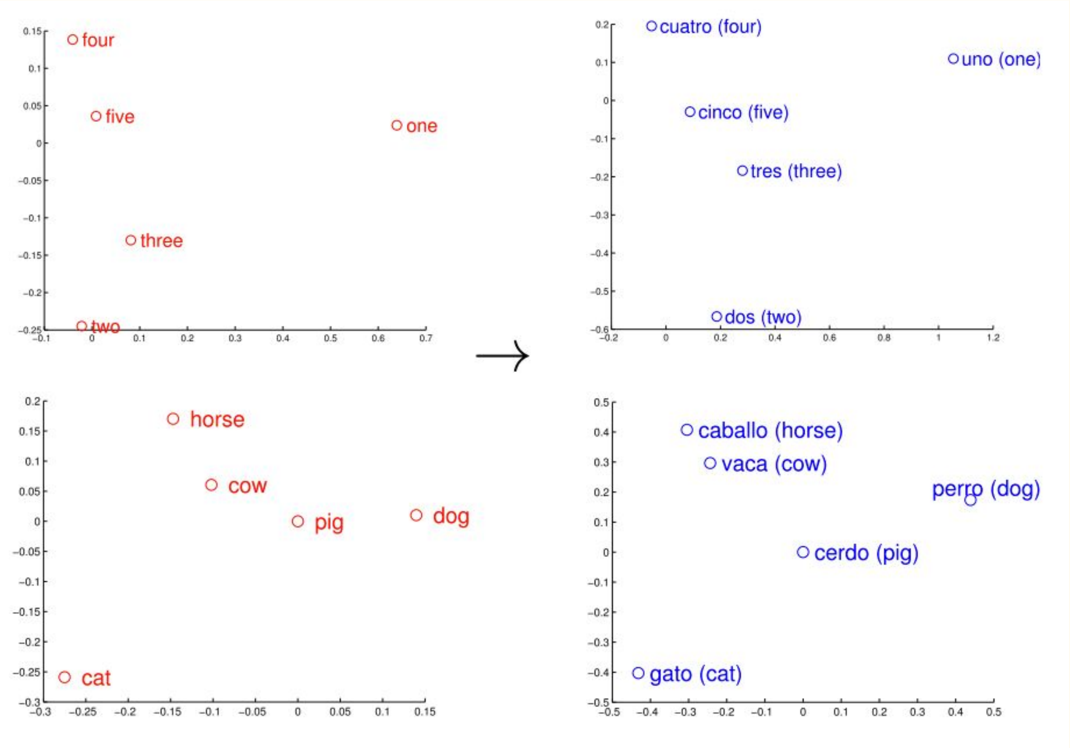
Mapping based approaches - Minimizing error

5000 most frequent words

Stochastic gradient to minimize d

$$\Omega_{\text{MSE}} = \sum_{i=1}^{n} \|\mathbf{W}\mathbf{x}_i^s - \mathbf{x}_i^t\|^2$$

$$\Omega_{\text{MSE}} = \|\mathbf{W}\mathbf{X}^s - \mathbf{X}^t\|_F^2$$



Mapping based approaches - CCA based mapping

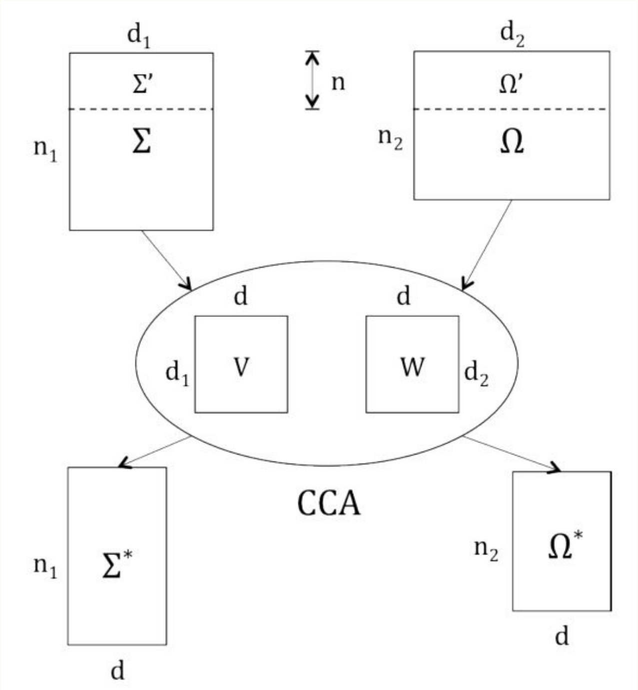
Mapping between languages

Canonical correlation analysis

1 matrix for every language

With 80% projection vectors with the highest corr

Separate synonyms and antonyms



Word level based on pseudo-bilingual corpus

No mapping

Xiao & Guo : seed bilingual dictionary > translate target > joint dictionary > random/or all switch

Center word switched during training > polysemy

Joint models - Bilingual language model

Joint formulation : $J = \mathcal{L}^s + \mathcal{L}^t + \Omega(s,t)$

Optimizing monolingual maximum likelihood objective of each language model with word alignment based regularization term

$$\mathcal{L} = -\log P(w_i \mid w_{i-C+1:i-1})$$

$$\Omega_s = \sum_{i=1}^{|V|^s} \frac{1}{2} \mathbf{x}_i^{s\top} (\mathbf{A}^{s\rightarrow t} \otimes \mathbf{I}) \mathbf{x}_i^s$$

Word level alignment with comparable data

Based on POS tag equivalence

Switch words with same POS

Take POS as context

Word level alignment with comparable data

Grounding language in images

Image as signal --> share the same cross lingual signal

Similarity between images --> interpolation with words

Also for audio signals

III Sentence level alignment models - Parallel data

Hard to obtain

Quality of the grain need some supervision

- Compositional sentence models
- Bilingual auto-encoder
- Bilingual skip-gram

Compositional sentence model

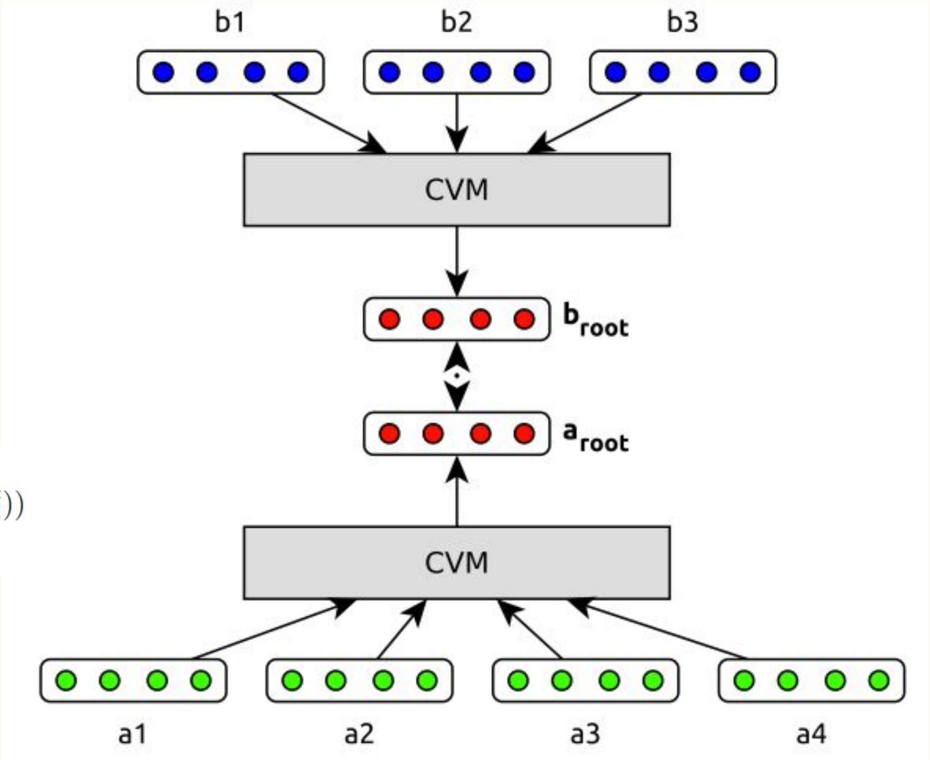
$$\mathbf{y}^s = \sum_{i=1}^{n} \mathbf{x}_i^s$$

$$\Omega = \frac{\lambda}{2}\|\mathbf{X}\|^2$$

$$E_{dist}(sent^s, sent^t) = \|\mathbf{y}^s - \mathbf{y}^t\|^2$$

$$\mathcal{L} = \sum_{(sent^s, sent^t) \in C} \sum_{i=1}^{k} \max(0, 1 + E_{dist}(sent^s, sent^t) - E_{dist}(sent^s, s_i^t))$$

$$J = \mathcal{L} + \Omega^s + \Omega^t$$



Bilingual autoencoder

Original source sentence reconstruction Target sentence

$$J = \mathcal{L}_{\text{AUTO}}^{s \to s} + \mathcal{L}_{\text{AUTO}}^{t \to t} + \mathcal{L}_{\text{AUTO}}^{s \to t} + \mathcal{L}_{\text{AUTO}}^{t \to s}$$

Encode a sentence as a sum of embeddings

Train autoencoder and decoder using softmax to reconstruct sentences and translations

Minimize the loss

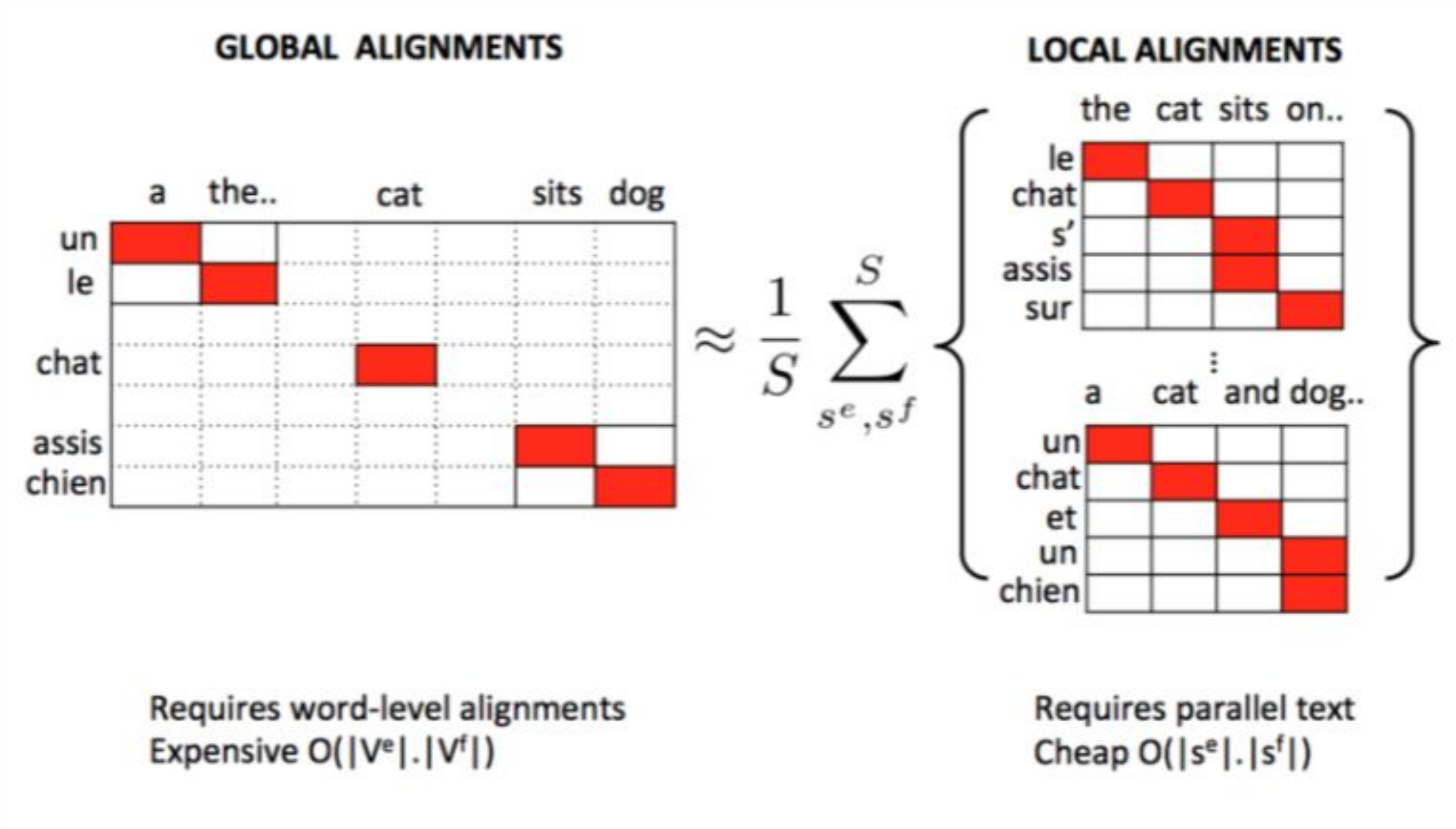
Bilingual skip-gram

Hypothesis : each word is aligned   Source <--> Target

Minimize the mean of the word representation

$$\Omega_{\text{BILBOWA}} = \| \frac{1}{m} \sum_{w_i^s \in sent^s}^{m} \mathbf{x}_i^s - \frac{1}{n} \sum_{w_j^t \in sent^t}^{n} \mathbf{x}_j^t \|^2$$

$$J = \mathcal{L}_{\text{SGNS}}^{s \to t} + \mathcal{L}_{\text{SGNS}}^{t \to s} + \Omega_{\text{BILBOWA}}$$


GLOBAL ALIGNMENTS
a
the..
cat
sits
dog
un
le
chat
assis
chien
LOCAL ALIGNMENTS
the
cat
sits
on..
le
chat
s'
assis
sur
a
cat
and dog..
un
chat
et
un
chien
$\approx \frac{1}{S} \sum_{s^e, s^f}^{S}$
Requires word-level alignments
Expensive O(|V^e|.|V^f|)
Requires parallel text
Cheap O(|s^e|.|s^f|)

Sentence alignment with comparable data

Grounding language in images

Images are used as pivot to induce a shared multimodal embedding space

Flickr 30k

Mix signal to obtain links

V Document level alignment models

- Approaches based on pseudo-bilingual document aligned corpora
- Concept-based methods
- Extensions of sentence-aligned models

Pseudo bilingual document aligned approaches

Merge and shuffle strategy

Concatenating the documents and shuffling them randomly

--> a strong and robust bilingual context for each word

--> completely random hence may be sub-optimal

Concept based models

Similarity if the same multilingual concept or topic is shared

$$\mathbf{x}_i^s = [P(w_1^s|w_i), \ldots, P(w_{|V^s|}^s|w_i), P(w_1^t|w_i) \ldots, P(w_{|V^t|}^t|w_i)]$$

Hypothesis : words share the same concept across language

--> Inversion of the index : not concept per wikipages but words per concepts

Extension of sentence-alignment models

Adjusting the regulation term based on the nature of the corpus

Regulazing word alignment and sentence in paragraph

$$\Omega = \alpha||\mathbf{y}_k^s - \mathbf{y}_k^t||^2 + (1-\alpha)\frac{1}{m}\sum_{w_i \in sent_k^s}^{m} \mathbf{x}_i^s - \frac{1}{n}\sum_{w_j^t \in sent_k^t}^{n} \mathbf{x}_j^t$$

$$J = \mathcal{L}_{\text{SGNS-P}}^s + \mathcal{L}_{\text{SGNS-P}}^t + \Omega$$

Extension of sentence-alignment models

To leverage data not sentence aligned but still aligned with something : Procrustes analysis

Learning monolingual representation of docs and align docs closely by transforming vector spaces

Consider docs as a bag of paragraph > alignment with paragraph

VI Evaluation

Tasks :

- How to measure the performance of algorithms ?
- Which metrics should be used ?
- How to compare results ?

Word similarity

How well results match with human selection ?

Word pairs are made and compare

SemEval 2017 introduced a cross similarity dataset

Can’t handle polysemy

Bilingual dictionary induction

Evaluation freely available

Manually constructed with non common spoken languages

Not 100% correct due to multiple translations but quiet accurate

Very good in specific domains if orthograph is taken into account...

Benchmarks

A website to evaluate word representation (from Faruqui and Dyer)

http://wordvectors.org

Focused mainly on evaluating monolingual word representation

Evaluation based only on word similarity dataset

Benchmark

Another website to evaluate word similarity, multiQVEC, bilingual induction, document classification

http://128.2.220.95/multilingual

Made by Ammar et al.

Monolingual and cross lingual word representation

Extension - Further works

Multi word expressions

Function words

Polysemy