Cross-lingual Embeddings

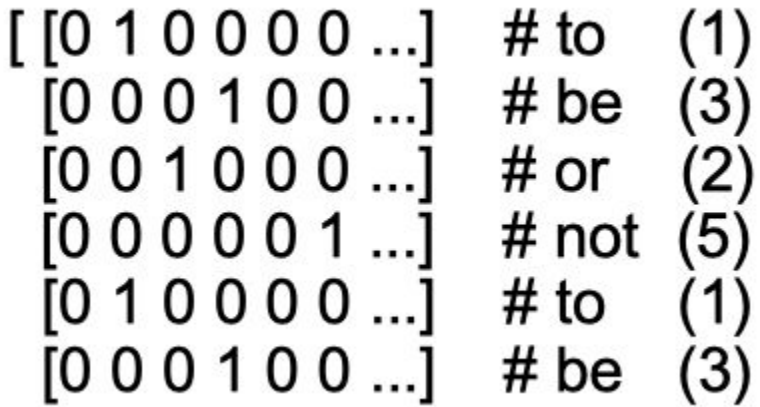
Deepen Naorem
Centre for Linguistic Science and Technology, IIT Guwahati
Under
Dr. Sanasam Ranbir Singh
&
Dr. Priyankoo Sarmah

CLST
IITG CENTRE FOR LINGUISTIC SCIENCE & TECHNOLOGY

Let us dive in to a brief history of word embedding

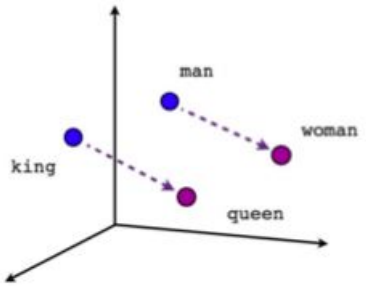
Representation of a word

- At the naive level
  - It can be a strings of letters.
    - Eg: “Election”
- One Hot-vector
  - An index in a vocabulary list.
  - Eg: Election=[0,0,1,0,0,0,0,.....,0]
  - Vector representation.
  - size= 1x|V|(as a row vector)
  - |V|=size of the vocabulary
- One Hot-vector is not an embedding.
  - Huge memory

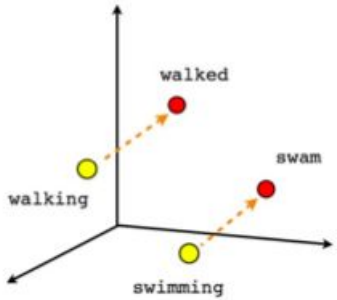


How to build better representation (Vector semantics)

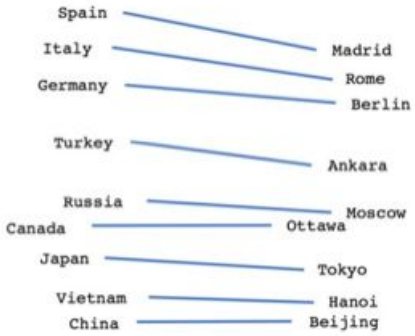
- The model Should deal with
  - Synonymy
  - Word Similarity
  - Word Relatedness
  - Connotation



Male-Female



Verb tense



Country-Capital

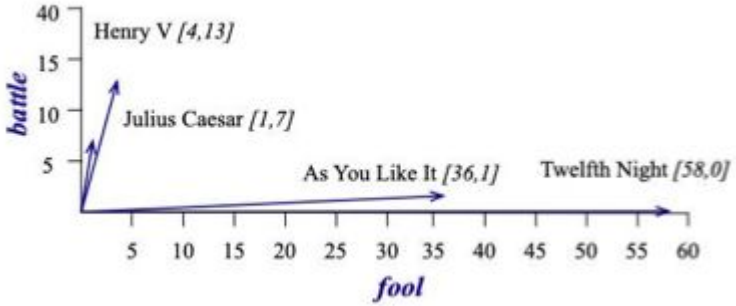
Importance of distributed representations

- Motivated by the distributional hypothesis
  - You shall know a word by the company it keeps.(Firth, 1957)
- Words with similar distributions will have similar meanings.
- Words that appear in similar contexts have similar meanings.
- Suppose you see these sentences:
  - Ong choi is delicious sautéed with garlic.
  - Ong choi is superb over rice
  - Ong choi leaves with salty sauces
  - spinach sautéed with garlic over rice
  - Chard stems and leaves are delicious
  - Collard greens and other salty leafy greens
- Ongchoi is a leafy green like spinach, chard, or collard greens.

Models from where embedding is derived

- Matrix based model like “count based model”.
  - Term-document matrix.
    - Each document is represented by a vector of words.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3



Curse of Dimensionality

- Reduce the size of this space from $\mathbb{R}^{|V|}$ to something smaller.
- Matrix Factorization
  - low-rank approximations of a massive matrix of word co-occurrence.
  - Methods
    - LSA (latent semantic analysis)
      - SVD matrix factorization
      - Word-document matrix. M_{ij} = # times word i appears in document j
    - PCA (principal component analysis)
      - PCA is a common method to factorize a matrix based on co-occurrence matrix.

Models from where embedding is derived(cont..)

- Prediction-based models
  - leverage language models, which predict the next word given its context.
- Traditional n-gram language model assumption:
  - The probability of a word depends only on context of n − 1 previous words”

$$\widehat{P}(s) = \prod_{i=1}^{k} P\left(w_i \mid w_{i-n+1} \dots w_{i-1}\right)$$

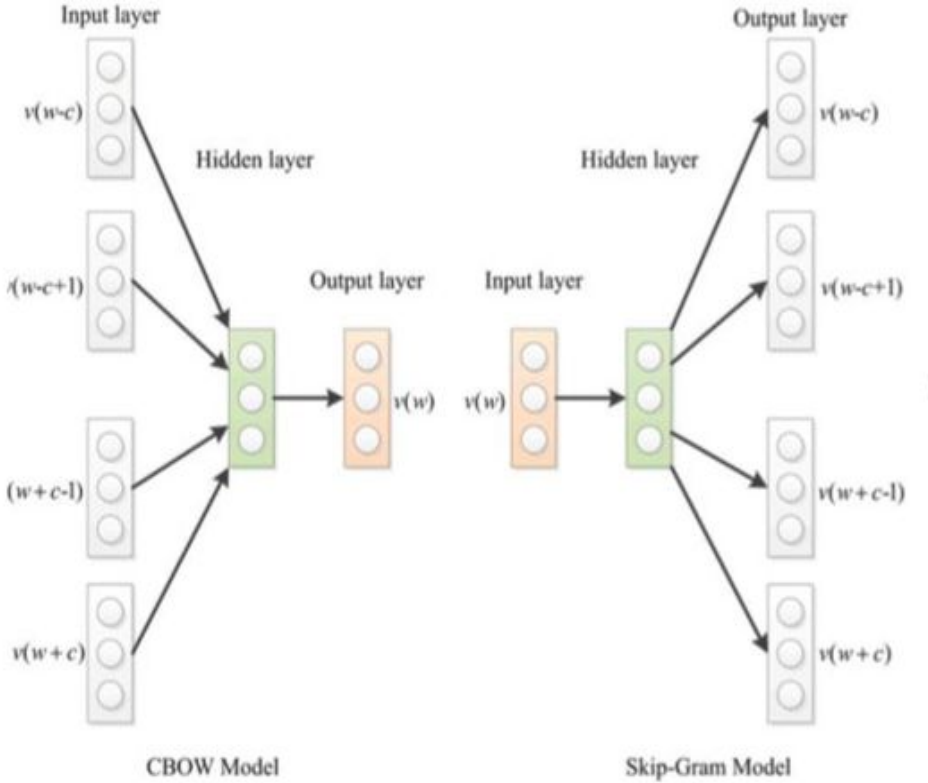
- Traditional methods are count-based; e.g., for trigrams:

$$p(w_3|w_1, w_2) = \frac{count(w_1,w_2,w_3)}{\sum_w count(w_1,w_2,w)}$$

- **Problems**: many sequences will have 0 probability

Word2vec - Skipgram & CBOW

- Neural Network Model
- CBOW: predict word given context
- Skip-gram: predict context given word



CBOW

is a group of related models that are used to produce word embeddings

So Finally

- . Formal Definition of “Word Embedding”
  - A distributed vector representation of a word.
  - A mapping from a one-hot-encoded space to a much lower dimensional continuous space.
- Bias in embeddings
  - But embedding analogies exhibit gender stereotypes.
    a. ‘man’ - ‘computer programmer’ + ‘woman’ = ‘homemaker’

Cross-Lingual Word embedding

- Representing lexical items from two different languages in a shared cross-lingual vector space.
  - Such that two similar lexical semantics words from two different languages are close to each other in shared vector space.
  - Lexical semantics looks at how the meaning of the lexical units correlates with the structure of the language.
  -

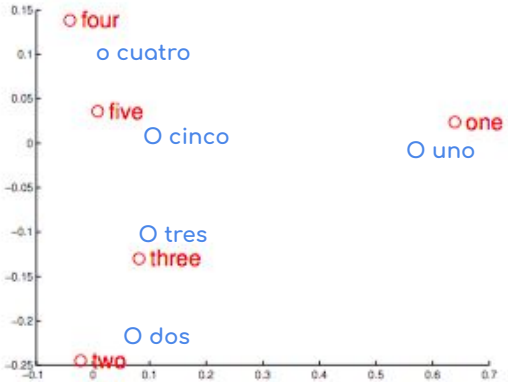
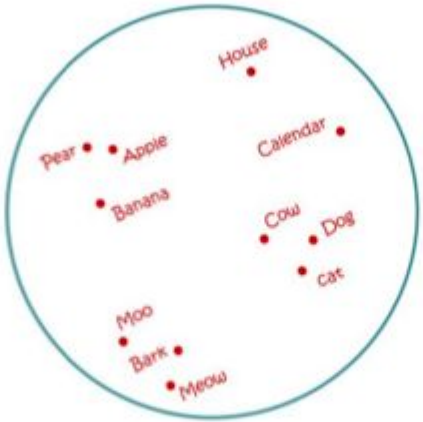


Fig : Shared embedding space
of English and Spanish

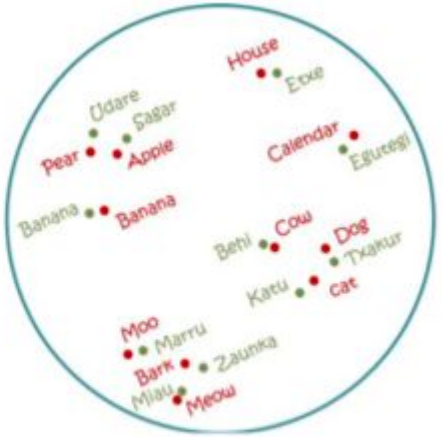
Why do we need Cross-lingual/Bi-lingual Embeddings?

- Bridge the language divergence
- Applications
  - Leverage the resource-richness of one language (e.g., English) in solving a problem in resource-constrained languages (e.g., Manipuri etc.)
  - Code-mix text

Embedding space for Spanish



Embedding space for English

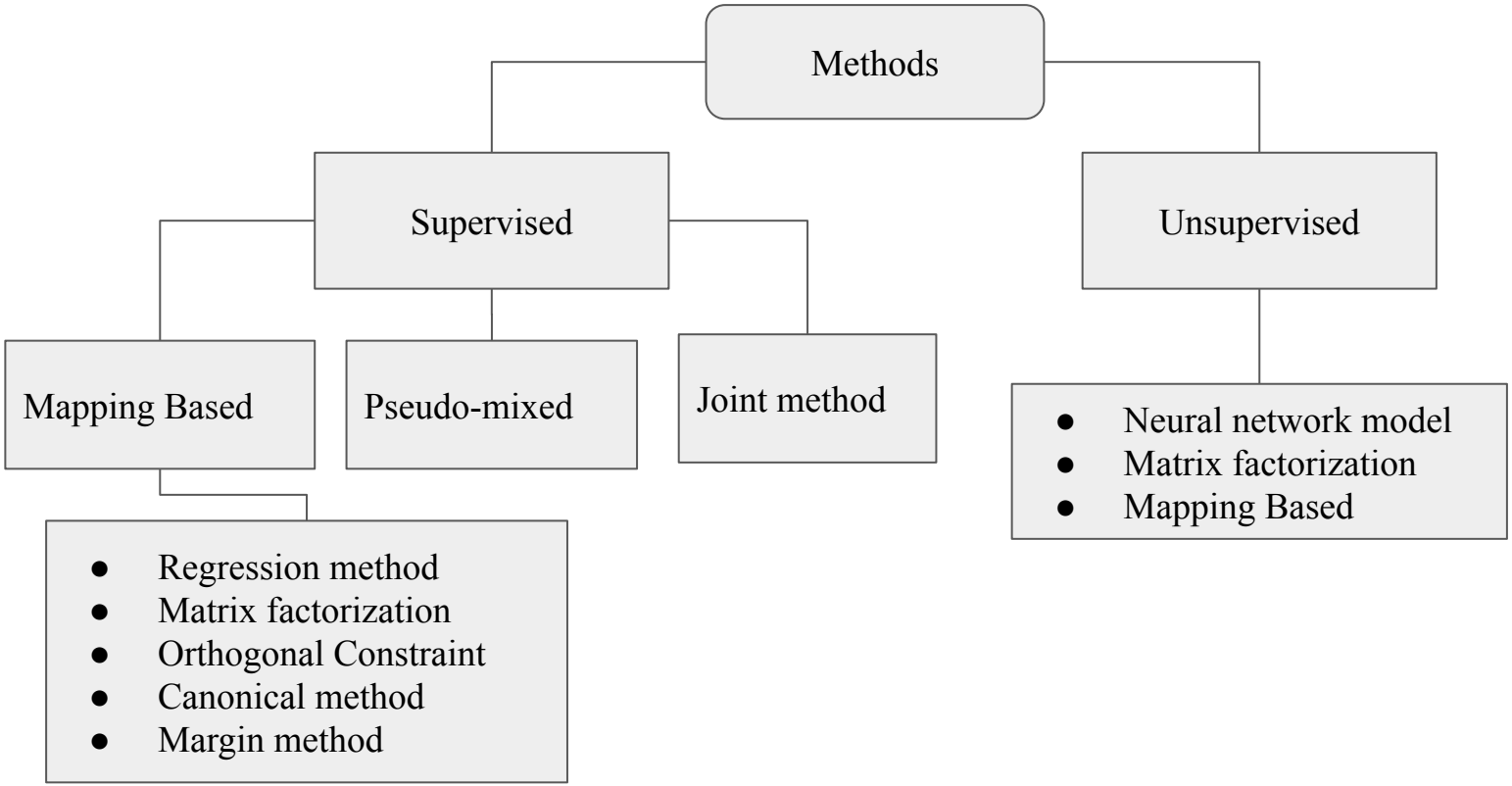


Shared Embedding space for Spanish and English

Applications

- Cross-lingual dependency parsing.
- Cross-lingual document classification.
- Machine translation.
- Cross lingual Name Entity and Part of Speech generation.

Structure of the state of the art method

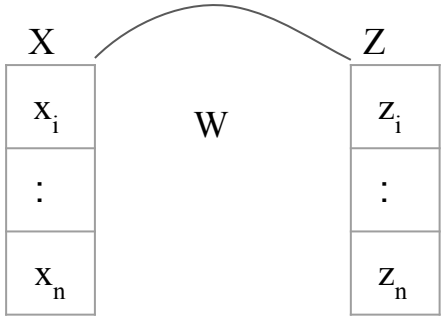


Supervised Method

- Need minimum two monolingual word embedding, say X and Z(Bilingual Setting)
- Monolingual word embeddings can be generated from parallel/non-parallel/comparable corpora.
- Bilingual Signal(or Bilingual dictionary) acts as supervision.

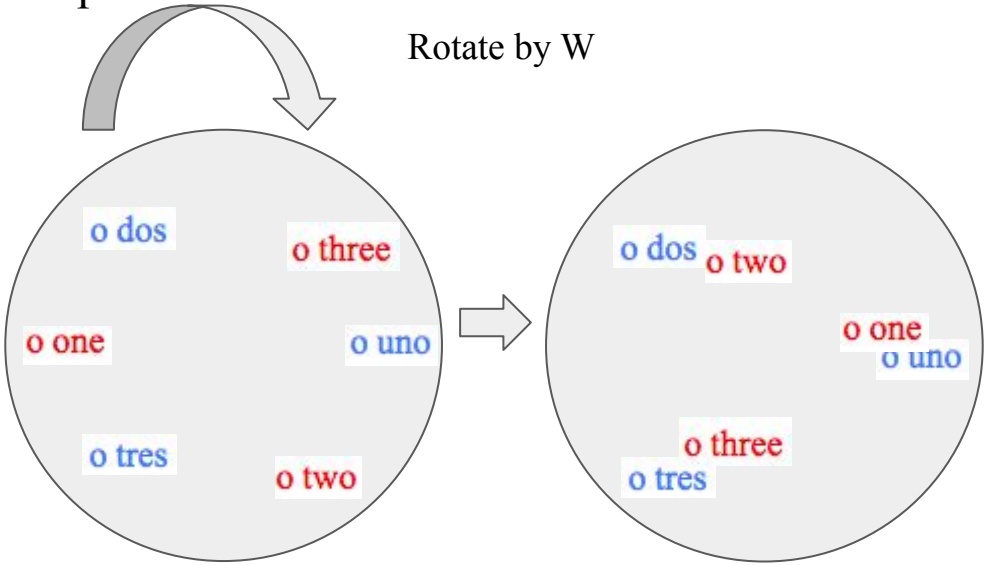
Supervised Method

Mapping Based method

- Required minimum two monolingual corpus(Bilingual setting)
- Bilingual dictionary(Alignment) to act as supervision.
- Learning a mapping function using Bilingual dictionary.
  - 

Learn a transformation matrix ‘ W’

New source word x_{new} will be translated as $z_{new}=Wx_{new}$

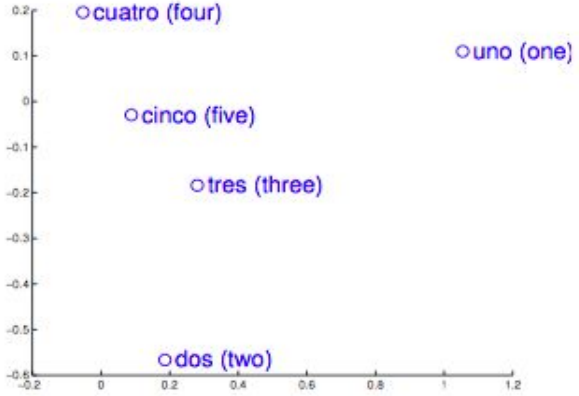


Supervised Method(Mapping based Method)

Regression Method

“Exploiting Similarities among Languages for Machine Translation”- Thomas Mikolov(2013)

- Mikolov et al.[1] used ridge regression to obtain the transformation matrix “W”

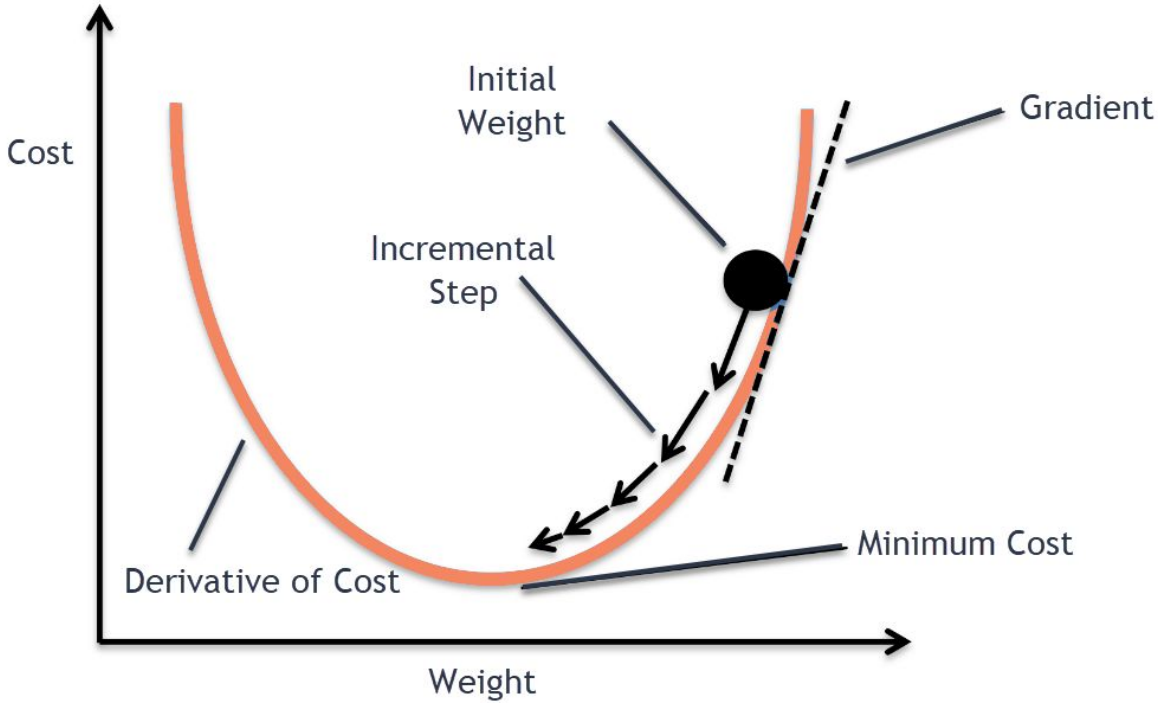


- Word pair (x_i,z_i) from the bi-lingual dictionary
- Find a transformation matrix W, given by

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

- New word embedding $z_{new}=Wx_{new}$

‘W’ can be solved through Stochastic Gradient descent, but it not linear with the size of the dictionary.



Supervised Method(Mapping based Method)

Orthogonal Constraint

“Normalized word embedding and orthogonal transform for bilingual word translation.”-Chao Xing(2015)

- Xing et al.[3] pointed out some inconsistencies among the objective functions of the embedding and the transform learning, as well as the distance measurement.
  - Proposes a solution which normalizes the word vectors on a hypersphere and constrains the linear transform as an orthogonal transform.

Skip gram model

$$P(w_{i+j}|w_i) = \frac{exp(c^T_{w_{i+j}} c_{w_i})}{\sum_w exp(c^T_w c_{w_i})}$$

The distance measure in the training is inner product

$$C^T_w c_{w'}$$

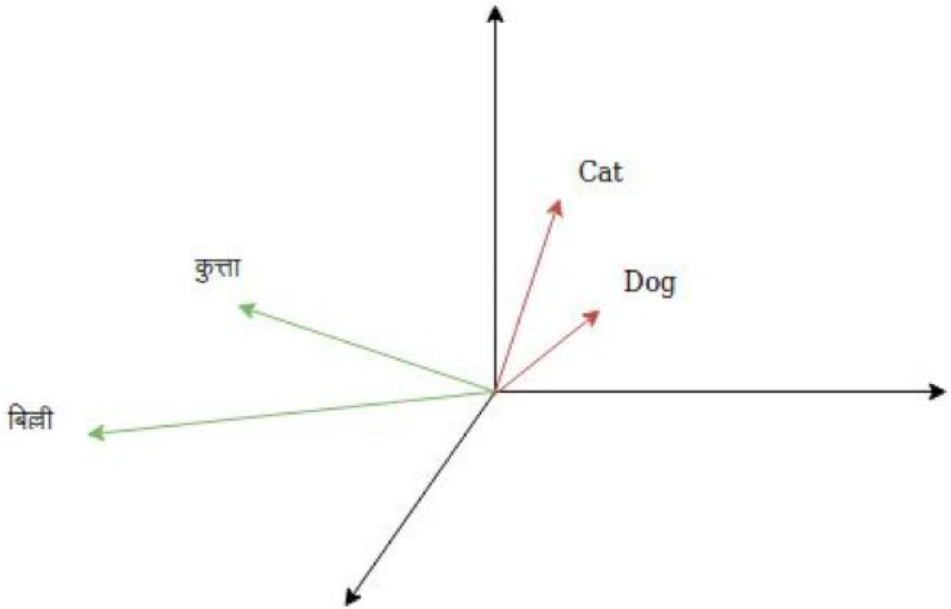
To estimate word similarity the matrix is often cosine similarity

$$\frac{c^T_w c_{w'}}{||c_w||\,||c_{w'}||}$$

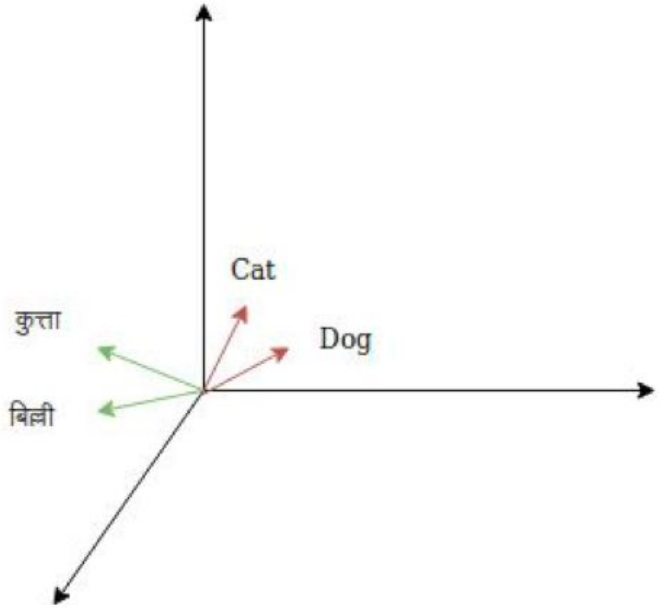
- How to solve the inconsistency in objective function of embedding?
  - Enforce the word vectors to be of unit length.(normalization)
  - The word vectors are located in a hyperspace.
  - The inner product falls back to cosine similarity.
- To normalised $x_i W$, ‘W’ has to be orthogonal matrix as orthogonal matrix preserve length of a vector.

Length normalization for maximum cosine

Visualization:



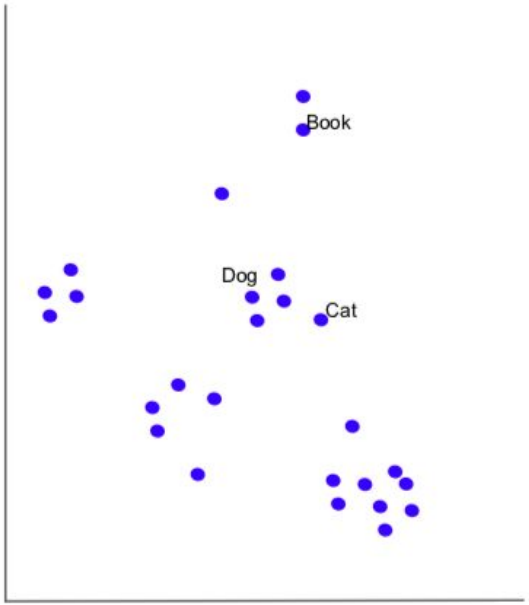
Unnormalized Embedding space

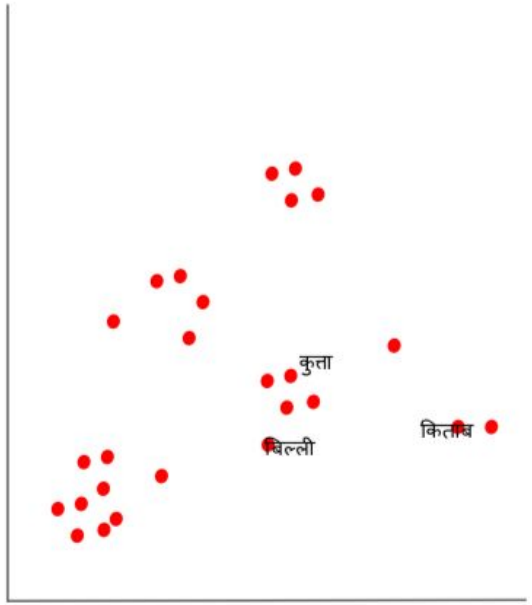


Normalized Embedding space

Why orthogonality is important

- Orthogonality is important to keep the monolingual property after transformation.
- Orthogonal transformation is length and angle preserving.
- Therefore it is an isometry of the Euclidean space (such as a rotation).





Procrustes problem

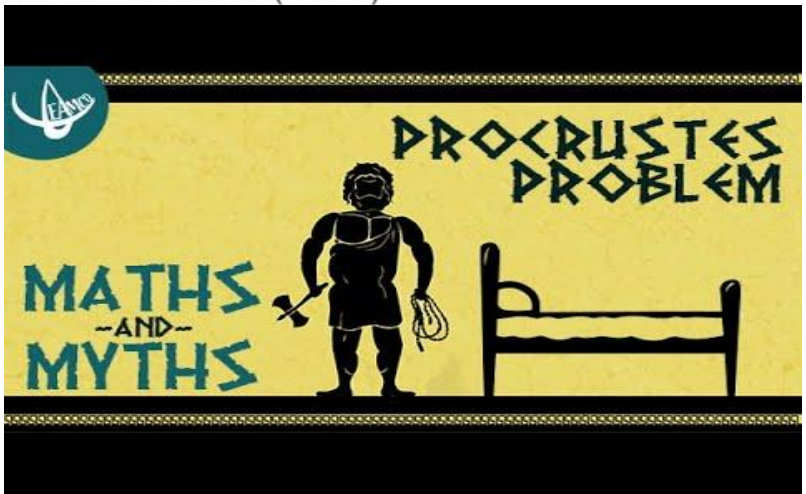
- A matrix approximation problem in linear algebra.
- Given two matrices X and Y the problem is to find a orthogonal matrix W which closely maps X to Y

$$\text{argmin}_W \, ||WX\text{-}Y|| \,, \qquad \text{subject to } WW^T = I$$

- This problem was originally solved by *Peter Schönemann* in his thesis (1964).

- Solution:

$$W = UV^T \text{ where } U\Sigma V^T = SVD(YX^T)$$



Monolingual invariance with orthogonal mapping

“Learning principled bilingual mappings of word embeddings while preserving monolingual invariance”-Mikel Artetxe(2016)

- Artetxe et al.[4] motivate orthogonality as a means to ensure monolingual invariance.
  - Experiments show that orthogonality is more relevant than length normalization, in contrast to Xing et al that introduce orthogonality only to ensure that unit length is preserved after mapping.

	EN-IT	Eng(Monolingual)
Original embeddings	-	76.66
Unconstrained mapping +length normalization	34.93 33.80	73.80 73.61
Constrained mapping +length normalization	36.73 36.87	76.66 76.66

Table: results in bilingual and monolingual tasks

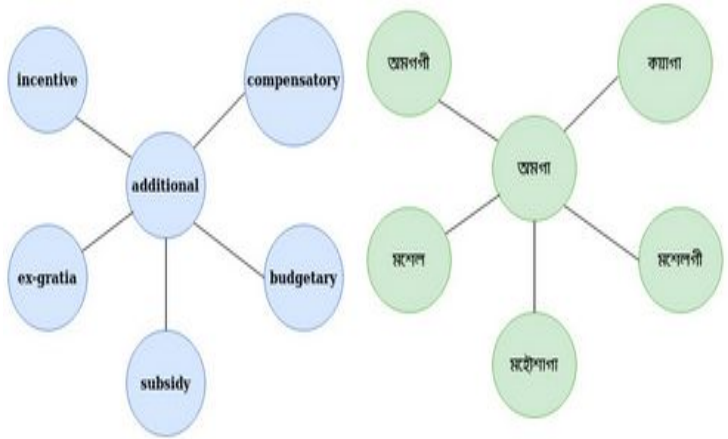
	EN-IT	Eng(Monolingual)
Original embeddings	-	76.66
Mikolov et al. Xing et al. Faruqui and Dyer et al	34.93 36.87 37.80	73.80 76.66 69.64
Proposed method	36.87	76.66

Table: Comparison with other work.

Cross-lingual embeddings in dissimilar language pairs

- Don’t work well in dissimilar language pairs
  - Non-isomorphic language pairs

	Unsupervised (Adversarial)	**Supervised** (Identical)	**Similarity** (Eigenvectors)
EN-ES	81.89	**82.62**	2.07
EN-ET	00.00	**31.45**	6.61
EN-FI	00.09	**28.01**	7.33
EN-EL	00.07	**42.96**	5.01
EN-HU	45.06	**46.56**	3.27
EN-PL	46.83	**52.63**	2.56
EN-TR	32.71	**39.22**	3.14
ET-FI	**29.62**	24.35	3.98



Sogard etal-2018

Cross-lingual embeddings in English-Manipuri

- Manipuri is a Tibeto-Burman languages
  - Widely spoken in the state of Manipur India.
  - Also a lingua franca among different communities in Manipur.
  - Two scripts:
    - Meitei Script or Meitei Mayek(ꯃꯩꯇꯩ ꯃꯌꯦꯛ □)
      - Pre Vaishnavism, reviving greatly for the last 10 years
    - Bengali Script
      - Post Vaishnavism

Cross-lingual embeddings in English-Manipuri

Semi-supervised	**En-Mn**		**En-It**	
	P@1	**P@5**	**P@1**	**P@5**
Artetxe 2017 [7] 25 dictionary	08.00	17.85	29.29	49.71
Artetxe 2017 [7]numerals	00.15	00.15	36.14	52.71
Unsupervised	**P@1**	**P@5**	**P@1**	**P@5**
Zhang 2017 [8]	00.00	00.00	30.45	44.53
Conneau 2017 [29]	00.00	00.00	46.45	57.25
Artetxe 2018 [18]	**04.46**	**12.15**	41.14	54.28
Biesialska 2020 [14]	00.23	01.78	42.23	58.76
Conneau 2017 [29]+iterative normalization	00.00	00.00	48.09	58.78
Supervised	**P@1**	**P@5**	**P@1**	**P@5**
Mikolov 2013 [4]	01.08	02.77	18.43	33.00
Artetxe 2016 [11]	12.15	23.08	46.00	65.29
Artetxe 2017 [7]	09.23	19.08	38.57	55.29
Iterative normalization [19]	**13.09**	**23.73**	**48.43**	**65.43**
MBERT [37]	00.14	00.14	11.00	15.28

Example

- Grouping of semantically similar words which are not direct translation

English word	nn1	nn2	nn3	nn4	nn5	Ground truth
stress	ফত্তবা (feel)	লায়নাশিং (diseases)	শিংনবশিং (grudges)	প্রোব্লেম (problem)	খুদোংথিবা (danger)	**স্ত্রেস**
maring	মরিং	রোংমৈ	আইমোল	কবুই	তাঙ্খুল	**মরিং**
philosopher	লম্বোইবা (saint)	ফিলোসোফর	কবি (poet)	শৈশকপা (singer)	স্বামি	**ফিলো সোফর**
thongkong	নাগামপাল	বাজারদগী	ককব্রা	অরোং	নাওরেমথোং	**থোং খোং**
pena	ফোল্ক	জগোই	আর্টিষ্ট	মাইবী	খুনুং	**পেনা**

Morphology as a challenge

- Manipuri is a morphologically rich language

English word	nn1	nn2	nn3	nn4	nn5	Ground truth
bungalow	সেক্রেতরিএত্তা	বঙ্গলোদা	রাজ	ভবন	চেম্বরদা	**বঙ্গলো**
fish	ঙাগী	কিহোম	পোৎথোক	শঙ্গোম	মচিঞ্জাক	**ঙা**
watching	য়েংবীয়ু	ইবদা	খঙঙি	নখোয়না	ঐবু	**য়েংলগা**
mentioned	পল্লহুনা	ফোঙদোকখিবা	এগ্রীমেন্তকী	পনখিবা	লিংবগী	**পনখি**
name	মিংদা	মমিংদা	ওপ্সন	ফোতো	একাউন্	**মমিং**


Gracias
Kiitos
Merci
Grazie
Vielen_Dank
ಧನ್ಯವಾದಗಳ
Thank_You

References

- **Distributional Structure (Zellig S. Harris)**
- **Word Embeddings: (A Survey Felipe Almeida Geraldo Xexeo)**
- **Word Embeddings through Hellinger PCA.**
- **Latent Semantic Analysis (LSA) (Deerwester et al.)**
- **NNLMs: A Neural Probabilistic Language Model, Bengio et al. (2003). Good blog post about this & new techniques (2017)**
- **Word2vec: ICLR paper with CBOW (2013); NIPS paper with Hierarchical Softmax and Negative Sampling (2013).**
- **Exploiting Similarities among Languages for Machine Translation- Thomas Mikolov(2013)**
- **Normalized word embedding and orthogonal transform for bilingual word translation- Chao Xing(2015)**
- **Learning principled bilingual mappings of word embeddings while preserving monolingual invariance- Mikel Artetxe(2016)**
- **OFFLINE BILINGUAL WORD VECTORS, ORTHOGONAL TRANSFORMATIONS AND THE INVERTED SOFTMAX -Samuel L. Smith(2017)**